



Identification et classification de composés reprotoxiques par des approches de toxicogénomique prédictive

Thomas Darde

► To cite this version:

Thomas Darde. Identification et classification de composés reprotoxiques par des approches de toxicogénomique prédictive. Médecine humaine et pathologie. Université de Rennes, 2017. Français. NNT : 2017REN1B022 . tel-01751766

HAL Id: tel-01751766

<https://theses.hal.science/tel-01751766>

Submitted on 29 Mar 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

THÈSE / UNIVERSITÉ DE RENNES 1

sous le sceau de l'Université Bretagne Loire

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Biologie et Sciences de la Santé

École doctorale Biologie-Santé

présentée par

Thomas DARDE

Préparée à l'IRSET Inserm U1085

Équipe « Physiology and physiopathology of the urogenital tract »
UFR Sciences de la Vie et de l'Environnement

**Identification et
classification de
composés
reprotoxiques par
des approches de
toxicogénomique
prédictive**

Thèse dirigée par :

Frédéric CHALMEL

Chargé de recherche - Inserm U 1085 IRSET

Antoine ROLLAND

Maître de conférence - Inserm U 1085 IRSET

et soutenue à Rennes le 3 octobre 2017

devant le jury composé de :

Karine AUDOUZE

Maitre de conférences – Université Paris Diderot
Rapporteur

Odile LECOMPTE

Maitre de conférences – iCube
Rapporteur

Farzad PAKDEL

Directeur de recherche – Inserm U1085 IRSET
Président du jury

Serge NEF

Professeur - University of Geneva Medical School
Examineur

Christian DELAMARCHE

Professeur – Université Rennes 1
Examineur

Frédéric CHALMEL

Chargé de recherche – Inserm U1085 IRSET
Directeur de thèse

Table des matières

Table des matières.....	ii
Liste des tableaux.....	vi
Liste des figures	viii
Liste des abréviations.....	x
Remerciements.....	xiv
Préambule	1
1. Introduction.....	3
1.1. Le système endocrinien et ses hormones	3
1.1.1. Qu'est-ce que le système endocrinien ?.....	3
1.1.2. Présentation des différents types d'hormones.....	5
1.1.3. Le mécanisme d'action des hormones	7
1.1.4. La régulation du système endocrinien	9
1.1.5. Les glandes endocrines et leurs hormones.....	11
1.2. Le testicule	19
1.2.1. Généralités sur l'anatomie du testicule adulte	19
1.2.2. Fonction exocrine du testicule	20
1.2.3. Fonction endocrine du testicule	25
1.2.4. Mécanismes de régulation de la stéroïdogénèse	28
1.3. Les perturbateurs endocriniens	31
1.3.1. Histoire de la perturbation endocrinienne.....	31
1.3.2. Un souci de définition.....	34
1.3.3. Mécanisme(s) d'action d'un perturbateur endocrinien.....	35
1.3.4. Effets des perturbateurs endocriniens sur l'Homme.....	36
1.3.5. Présentation de perturbateurs endocriniens connus	46
1.4. Évaluation des risques.....	53
1.4.1. La réglementation européenne.....	53
1.4.2. Réglementation non adaptée aux PEs?	56

1.4.3.	Outils de détection de toxicités	58
1.4.4.	Tests spécifiques de détection des perturbateurs endocriniens.....	61
1.5.	La toxicologie prédictive	67
1.5.1.	Outils de toxicologie prédictive	67
1.5.2.	Méthodes de prédiction.....	73
2.	Objectifs de ma thèse	91
3.	Matériels et méthodes	93
3.1.	Ressources informatiques et environnement de travail	93
3.1.1.	Langages de programmation.....	93
3.1.2.	Environnement web	96
3.2.	Outils et ressources bio-informatiques.....	101
3.2.1.	Suites logicielles	101
3.2.2.	Navigateurs de génome.....	101
3.2.3.	Galaxy	102
3.2.4.	Vocabulaires contrôlés et ontologies	104
3.2.5.	Bases de données	105
3.2.6.	Description des différents types de fichiers	107
3.3.	Estimateurs mathématiques	111
3.3.1.	Estimateurs de distances	111
3.3.2.	Évaluation d'un modèle statistique de classification	113
3.4.	Méthodes statistiques	117
3.4.1.	Analyse en composantes principales	117
3.4.2.	Enrichissement – Loi hypergéométrique	122
3.4.3.	Méthodes de prédiction de classe	123
3.5.	Méthodes appliquées dans le cadre du projet ChemPSy	127
3.5.1.	Méthodes appliquées sur les données de la CTD.....	127
3.5.2.	Méthodes appliquées sur les données issues de GEO.....	129
4.	Résultats et discussion	137
4.1.	ChemPSy : un système de priorisation pour les substances chimiques	137
4.1.1.	Résultats obtenus à partir de la CTD	137

4.1.2.	Résultats obtenus à partir de GEO	140
4.1.3.	Conclusion et perspectives du projet ChemPSy	159
4.2.	TOXsIgN : un espace de dépôt pour les signatures toxicogénomique	163
4.2.1.	Manuscrit en préparation	165
4.2.2.	Résultats et discussion	191
4.2.3.	Conclusion et perspectives.....	196
4.3.	The ReproGenomics Viewer (RGV) : un navigateur de génome pour les données de reprogénomique	199
4.3.1.	Publication du ReproGenomics Viewer.....	199
4.3.2.	Résultats	213
4.3.3.	Discussion et perspectives	217
5.	Autre publication	219
	Conclusion	231
	Références.....	235
	Annexes.....	ii
1.	Conditions expérimentales des groupes sélectionnés	ii
1.1.	Conditions expérimentales du groupe 1	ii
1.2.	Conditions expérimentales du groupe 2.....	iv
1.3.	Conditions expérimentales du groupe 3.....	v
	Résumé.....	

Liste des tableaux

Tableau I.	Tableau récapitulatif des différentes hormones et leurs actions	16
Tableau II.	Tests officiels de l'OCDE pour la détection de perturbateurs endocriniens.....	62
Tableau III.	Symboles utilisés dans le moteur de recherche Elasticsearch	99
Tableau IV.	Colonnes composant un fichier BED.....	108
Tableau V.	Colonnes composant un fichier GFF	109
Tableau VI.	Paramètres adaptés aux matrices discrétisées de ChemPSy	112
Tableau VII.	Matrice de confusion.....	114
Tableau VIII.	Tableau de confusion	114
Tableau IX.	Résumé des données intégrées dans ChemPSy	140
Tableau X.	Corrélations entre les estimateurs mathématiques de similarité toxicogénomique et d'effets toxicologiques.....	142
Tableau XI.	Résultat de la classification des CE sans récursivité.....	144
Tableau XII.	Résultat de la classification des CE avec récursivité	145
Tableau XIII.	Molécules du premier groupe sélectionné.	148
Tableau XIV.	Molécules présentes dans le second groupe.....	150
Tableau XV.	Molécules présentes dans le troisième groupe.....	152
Tableau XVI.	Spécificité et sensibilité moyenne pour chaque technique de classification	155
Tableau XVII.	Liste des études disponibles dans RGV	214

Liste des figures

Figure 1.	Modes de communications cellulaires.....	4
Figure 2.	Schéma des différents types d'hormones.....	6
Figure 3.	Mécanismes d'action des hormones	8
Figure 4.	Schéma du système endocrinien	10
Figure 5.	Schéma de l'hypothalamus et de l'hypophyse.....	12
Figure 6.	Anatomie du testicule adulte humain.....	20
Figure 7.	Coupe d'un tubule séminifère.....	21
Figure 8.	Fonction exocrine du testicule : la spermatogénèse.....	24
Figure 9.	La stéroïdogénèse.....	27
Figure 10.	Mécanisme de régulation de la stéroïdogénèse.....	29
Figure 11.	Historique de la perturbation endocrinienne.....	33
Figure 12.	Différents modes d'action des PEs	37
Figure 13.	Schéma du syndrome de dysgénésie testiculaire selon Skakkebaek <i>et al.</i> 2016 42	
Figure 14.	Schéma représentatif de la cryptorchidie et de l'hypospadias.....	45
Figure 15.	Planning et processus d'évaluation des risques chimiques selon REACH.....	54
Figure 16.	Représentation d'une courbe dose-réponse et calcul de la DJA	57
Figure 17.	La toxicologie computationnelle et les bases de données.....	69
Figure 18.	Génération de modèles prédictifs.....	72
Figure 19.	Création et utilisation des QSARs	75
Figure 20.	Calcul de distance topologique au sein de la molécule de cholestérol	77
Figure 21.	Diagramme de <i>flow</i> de modèle PBPK	81
Figure 22.	Représentation des modèles PBPK les plus utilisés	84
Figure 23.	Schématisation du modèle MVC.	96
Figure 24.	Exemple de workflow sous Galaxy.	103
Figure 25.	Représentation en deux dimensions de l'ACP : matrice de coordonnées et de poids. 119	
Figure 26.	Tableau disjonctif complet et tableau de Burt	121
Figure 27.	Description des ensembles de la loi hypergéométrique.....	122

Figure 28.	Discrimination des données par SVM	124
Figure 29.	Travail effectué sur les données obtenues à partir de la CTD	128
Figure 30.	Représentation d'une puce à ADN	131
Figure 31.	Réduction des matrices	132
Figure 32.	Répartition des composés en fonction des 111 groupes en pourcentage cumulé 138	
Figure 33.	Exemple de courbe de classification d'une CE.....	146
Figure 34.	Estimation de la meilleure distance de classification	147
Figure 35.	Classification du fluconazole à 394 mg/kg, 6 heures	149
Figure 36.	Courbe de classification du fénopropène à 52 mg/kg, 1 jour	151
Figure 37.	Courbe d'erreur cumulée de la reclassification	156
Figure 38.	Courbe de pourcentage de réussite cumulé de la reclassification.....	157
Figure 39.	Capture d'écran de l'outil de recherche avancée de TOXsIgN	194
Figure 40.	Capture d'écran du résultat de comparaison (top 10) de la signature du fluconazole	195
Figure 41.	Outil de conversion de RGV	216

Liste des abréviations

ACM. : Analyse des correspondances multiples
ACP : Analyse en composantes principales
ADME : Absorption, Distribution, Métabolisme et Excrétion
AINS : Anti-inflammatoire non stéroïdien
AR : Récepteur à androgènes
BPA : Bisphénol A
CAS : Chemical Abstracts Service
CE : Condition Expérimentale
ChemPSy : Chemical Prioritization System
DES : Diethylstilbestrol
DDT : Dichlorodiphényltrichloroéthane
ECHA : European Chemicals Agency
EINECS : European INventory of Existing Commercial chemical Substances
OMS : Organisation Mondiale de la Santé
PBPK : Physiologically Based Pharmacokinetic
PCB : Polychlorobiphényle
PCOS : PolyCystic Ovary Syndrome
PE : Perturbateur Endocrinien
PLS : Partial Least Squares
POP : Polluant Organique Persistant
QSAR : Quantitative Structure-Activity Relationship
QSRP : Quantitative Structure-Property Relationship
RA : Read Across
REACH : Registration, Evaluation, Authorization and Restriction of Chemicals
RGV : The ReproGenomics Viewer
SE. : Système Endocrinien
SVM : Support Vector Machine
TOXsIgN : TOXicogenomic sIgNatures
US EPA : United States Environmental Protection Agency

Dans la vie, j'avais deux ennemis : le vocabulaire et les épinards. Maintenant j'ai la botte secrète et je bouffe plus d'épinards. Merci, de rien, au revoir messieurs-dames.

Franck Pitiot

Remerciements

Merci au Docteur et Professeur **Bernard Jégou**, directeur de l'IRSET, directeur de la Recherche et de l'innovation à l'École des Hautes Études en Santé Publique, pour son accueil au sein de l'IRSET. Votre culture scientifique m'impressionnera toujours.

Merci au Docteur **Nathalie Dejucq-Rainsford**, directrice de l'équipe de recherche « Environnement viral et chimique, & Reproduction », de m'avoir accueilli au sein de son équipe ainsi que pour les conseils scientifiques dispensés lors des réunions d'équipe.

Merci à **Olivier Collin**, responsable de la plateforme bio-informatique GenOuest, de m'avoir accepté au sein de sa plateforme.

Je tiens à adresser mes sincères remerciements au Docteur **Karine Audouze**, maître de conférences à l'Université Paris Diderot ainsi qu'au Docteur **Odile Lecompte**, maître de conférences à l'Université de Strasbourg, pour avoir accepté d'être rapporteur de cette thèse.

Mes remerciements vont également au Docteur **Farzad Pakdel**, directeur de recherche à l'IRSET et au Professeur **Christian Delamarche**, maître de conférences à l'Université de Rennes 1 pour avoir accepté d'examiner ce travail de thèse.

Je tiens à adresser mes profonds remerciements aux Docteurs **Frédéric Chalmel** et **Antoine Rolland**, respectivement chargé de recherche à l'IRSET et maître de conférences à l'Université de Rennes 1, tous deux directeurs de cette thèse. Un grand merci d'avoir cru en moi, vous m'avez offert le cadre idéal pour réaliser cette thèse.

À mon tuteur de thèse, le Docteur **Jean-Jacques Lareyre** et les membres de mon comité de thèse, les Docteurs **Marie-Agnès Coutellec** et **Christophe Hitte**, mes sincères remerciements pour vos précieux conseils, critiques et échanges durant ces trois dernières années.

Plus généralement, je tiens à remercier les trois équipes avec qui j'ai pu apprendre, échanger, rire et surtout m'épanouir. Merci donc aux membres de **l'équipe 8 de l'IRSET**, à ceux de l'équipe **Symbiose de l'INRIA** et également merci aux membres de la **plateforme GenOuest**.

Durant ces trois années, j'ai eu la possibilité de travailler auprès de personnes qui m'ont permis d'avancer dans mes projets en me « prêtant » leurs connaissances et compétences. Ainsi je remercie sincèrement Séverine, Laurianne et Christelle pour leur travail dans le cadre de la validation des composés de ChemPSy. Merci à Sébastien pour sa patience et son efficacité quand il s'agit de gérer le

cluster de l'IRSET. Enfin merci à Catherine et Véronique d'avoir réussi à simplifier ce qui ne l'est pas : les démarches administratives.

Parce que la thèse c'est aussi s'impliquer à différent niveau dans des associations, encadrements ou réseaux. Je tiens à remercier les membres de l'UPSET, Julie, Julie (pas Julie, mais l'autre), Charly et Laetitia pour votre bonne humeur et votre motivation. On s'en souviendra de ces journées « Chercheurs d'aujourd'hui et de demain » et « Jeunes chercheurs de l'IRSET ». Merci à Pierre de m'avoir fait entrer au sein du NYRA (ex INYRMF). J'en profite ainsi pour remercier tous les membres du NYRA et plus particulièrement Dorte, Judit, Alexandra et Yoni, c'est un vrai plaisir d'avoir fait votre rencontre et ce fut un réel plaisir d'organiser le 9^e meeting à Rennes. Enfin merci aux stagiaires que j'ai encadrés, vous m'avez fait comprendre à quel point la lecture d'un rapport pouvait être difficile, mais également l'importance de se relire après avoir écrit un paragraphe.

La thèse ne se réalisant pas seul, je tiens à remercier toutes les personnes ayant participé de près comme de loin à ce travail en me soutenant, me relisant ou même en me changeant les idées.

Merci à Yvan, François et Jennifer, l'équipe EnginesOn. J'ai vraiment apprécié les sessions développement, brainstorming et pizzas chez François. Malheureusement je ne suis pas allé au bout de l'aventure avec vous, mais je serais toujours présent pour vous filer un peu coup de main au besoin. Vous m'avez beaucoup appris et je vous en remercie. Ma fibre entrepreneuriale renaîtra de nouveau.

Les mots me manquent pour te remercier Olivier O. Tu fais vraiment partie d'une espèce rare d'homme ours, mi-dieu du développement, mi-café. Sans ton aide j'aurais sans doute fini par devenir chauve à force de m'arracher les cheveux de la tête. Encore merci. Je n'oublie pas non Cyrile, Anthony et Olivier C. Merci à vous pour votre patience et vos explications. Je garderai en tête cette mémorable raclette GenOuest et l'assiette de 2,3kg d'Yvan.

Remi, je suis vraiment heureux d'être, un beau matin, passé dans ton bureau pour discuter d'un projet bancal. J'ai ainsi pu faire connaissance une personne géniale et accessoirement le seul homme sage-femme qu'il m'ait été donné de rencontrer. La porte du bureau est toujours ouverte pour discuter rhum.

Sans distraction, j'aurai sûrement perdu la tête. Je tiens donc à remercier le groupe de Capoeira Brasil pour m'avoir initié aux joies du papillon. Merci également aux potes Geeks, Quentin, Maxime et Ben Ben. Promis, on se refait de nouvelles soirées bientôt. Merci au groupe de « coupains », Nathalie, François, Antoine, Émilie, Anaïs, CamCam (aka. Binôme) et Quentin (remercié deux fois) pour votre soutien de près comme de loin.

Victo, voici maintenant plus de 4 ans qu’au moins une fois par semaine on discute de tout et n’importe quoi. Tu es vraiment une personne unique. Quoi que dise les gens ne change pas.

Sur un volet beaucoup plus personnel, je voudrais ici remercier du plus profond de moi ma famille pour leur soutien sans limites. Vous avez toujours été derrière moi pour me pousser à toujours faire mieux. Merci Moman, Tonton, Tata, Agnès, Fredo, Bro, Séverine, Maxou, Alain, Djamila, Mamie et Grand père. Tanguy, je sais que je n’ai pas l’air d’un scientifique, mais j’espère que tu changeras d’avis en lisant ce manuscrit. À Bastoune, ceci est ma façon de « jouer avec le game ».

Ma très chère voisine, elle est bien loin l’époque où je me nourrissais uniquement de céréales. J’ai commencé après toi, et je finis avant toi. La socio aura ta peau. Tu es sur la dernière ligne droite courage on fêtera ça ensemble.

Toujours sur le registre personnel, je tiens à adresser quelques mots aux personnes avec qui je travaille depuis maintenant 3 ans.

Fred, je t’ai rencontré pour la première fois lors de mon stage de M1 et dès lors tu n’as cessé de tout faire pour valoriser mes compétences. Depuis maintenant 5 ans tu es devenu beaucoup plus qu’un simple encadrant de stage ou de directeur de thèse. Tu es pour moi un mentor à la soif d’apprendre intarissable, quelqu’un d’entier qui ne fait jamais les choses à moitié et dont la passion pour son travail donne vraiment envie de s’investir à 200% pour être à la hauteur de tes espérances. Enfin au cours de ces dernières années tu es également devenu un très bon ami pour moi. Je n’oublierai pas les soirées zombicide, retro-gaming et les journées de surf. Le mot **MERCI** est encore trop faible pour exprimer ma gratitude, voilà pourquoi je le mets en gras.

Antoine (aka. Toto), comme pour Fred, tu représentes plus qu’un simple directeur de thèse pour moi. Tes critiques constructives et bienveillantes m’ont permis d’apprendre beaucoup. Tu es toujours là quand il s’agit d’aider ou d’échanger sur le surf et le beau temps. Avec Fred, vous formez un duo atypique, j’ai été fier d’avoir été encadré par vous deux. D’habitude je me fiche pas mal de ce que pensent les gens de moi, mais j’espère du fond du cœur de ne vous avoir jamais déçus au cours de ces 3 années et si un jour j’arrive à acquérir ne serait-ce que le quart de vos connaissances, je m’estimerai heureux.

Fred, Toto, encore **MERCI**.

Bertrand, maître du fractionné, gourou de la nage et de la pipette. Merci pour ta bonne humeur, il suffit de t’entendre rire pour savoir que l’on va passer une bonne journée. Tes conseils pour m’améliorer à la course commencent enfin à porter leur fruit. J’ai hâte de faire le MUD DAY à tes côtés.

Emmanuelle, merci infiniment pour tes conseils et les discussions animées sur comment recoder R, quel est le meilleur test statistique ou encore si le fortran allait revenir à la mode. J'ai appris beaucoup grâce à toi.

Aurélié, ma partenaire de course. Heureusement que ton assiduité est là, sinon il y a fort à parier que je ne me serais jamais levé pour aller courir le dimanche avec toi. Je te remercie pour tout.

Clément, j'espère que ces 6 mois de stage ne t'auront pas trop traumatisé. Tu as pu assister à la partie la plus difficile de la thèse. Si même ça ne t'as pas rebuté alors fonce, tu es bourré de compétences (malgré ton caractère d'ours mal léché).

Enfin pour finir trois personnes ont marqué au fer rouge cette thèse.

Angélique, tu es la première. Même si cela ne fait pas longtemps que tu es parmi nous, ta sociabilité, ta bonne humeur et ta joie de vivre font que j'ai l'impression que tu es là depuis le début. Cela vient sans doute aussi de nos nombreuses conversations sur tout et n'importe quoi (surtout n'importe quoi). Merci pour tous ces bons instants passés en ta compagnie.

Dude, mon seul regret est de ne pas t'avoir connu plus tôt. On a formé une sacrée équipe ensemble transcendant le copinage thésard-thésard pour devenir de vrais amis. Toujours là pour me dispenser de ton savoir de pharmacien. La semaine à Florence, celle de l'ETW, les soirées geeks ou encore les lundis GOT font partie de mes meilleurs souvenirs de thèse.

Princesse Marcel... Que dire ? Voilà maintenant 4 ans que nous sommes compagnons d'infortune, partageant nos réussites et nos moments de doutes. Cette rencontre avec toi, au détour d'une salle stagiaire surchauffée, près de la table de ping-pong, fait partie de mes plus belles rencontres. Tu es la prochaine alors courage. Quoiqu'il arrive si tu as besoin de moi je serai toujours là pour te filer un coup de main.

Enfin pour finir, je tiens à remercier celle qui a su me supporter, me consoler, me motiver, me recadrer et j'en passe, Maëlle. Je sais que je ne suis pas facile quand le manque de sommeil et le stress se fait ressentir. Merci pour tout. Tu œuvres depuis maintenant plus de 10 ans à faire de moi une meilleure personne, et j'espère que cela continuera pour au moins 10 années de plus. Encore **MERCI**.

Préambule

De nos jours, la production mondiale de substances chimiques, dont certaines d'entre elles s'avèrent nocives pour l'environnement et la santé humaine, est 400 fois plus importante qu'il y a un siècle. Selon l'Inventaire Européen des Substances Chimiques Commerciales Existantes (EINECS), le nombre de produits chimiques manufacturés dans le monde est estimé à plus de 100.000, dont seulement 3% ont fait l'objet d'analyses approfondies pour en évaluer leur toxicité et établir des liens avec des pathologies et des phénotypes délétères chez l'Homme. Ce chiffre ne semble cependant représenter qu'une infime partie du nombre total de substances chimiques présentes dans le monde, revendiquées à plus de 140 millions par le site du CAS (Chemical Abstract Service) de l'American Chemical Society.

Dans son livre « Les perturbateurs endocriniens », le Professeur Olivier Kah évoque « l'invasion incontrôlée » de ces composés chimiques (Kah 2016). Ces derniers sont omniprésents, rendant notre exposition quasi-constante aux pesticides (McKinlay et al. 2008), plastifiants (Romani et al. 2013), cigarettes, cosmétiques, détergents, médicaments, et bien d'autres. Il est aujourd'hui très difficile de mesurer l'exposition d'un individu à un composé précis tant les sources d'exposition sont multiples. Ce problème d'évaluation est d'autant plus préoccupant, que si aujourd'hui nous sommes dans l'impossibilité de définir ce à quoi nous sommes réellement exposés, qu'en est-il de l'exposition fœtale chez la femme enceinte, et quelles peuvent être ses conséquences pour l'enfant à naître ?

Les effets des polluants et des produits chimiques sur la santé humaine et animale sont depuis plusieurs années au centre des préoccupations scientifiques et politiques. Depuis le début des années 1990, la prise de conscience sur la présence de ces substances dans notre environnement a mené les autorités sanitaires à mettre en place de nombreux programmes de gestion et d'évaluation des risques, tels que la législation REACH (Registration, Evaluation, Authorization of Chemicals). Parmi ces composés chimiques, la famille des perturbateurs endocriniens (PE) est au centre des débats. En effet, il est possible de se rendre compte de l'ampleur de la polémique en entrant les mots « perturbateurs endocriniens » dans un moteur de recherche. Le nombre d'articles à destination du grand public a conduit à une prise de conscience générale sur l'impact néfaste de ces substances sur l'Homme. D'origine naturelle ou synthétique, les PEs peuvent altérer le fonctionnement du système hormonal animal. Certaines de ces substances pourraient induire des effets néfastes sur la santé d'un individu, de sa descendance ou d'une population et ainsi constituer un véritable danger sanitaire. L'intérêt pour ce sujet croît régulièrement et en parallèle, des plateformes de décision réunissant des représentants de la science, de l'industrie et des autorités ont été créées comme par exemple le programme de détection des PEs

lancé en 1996 par l'US EPA (*Environmental Protection Agency*) (Usepa 1998), le plan national santé-environnement (PNSE) débuté en 2004 ou encore le pôle national applicatif en toxicologie et écotoxicologie lancé en 2009 en France. Elles prévoient des mesures à court, moyen et long terme, parmi lesquelles l'identification des PE, la priorisation de l'évaluation des risques, l'élaboration de programmes de monitoring, la recherche sur les mécanismes d'action de ces substances et enfin l'adaptation de la législation. Depuis 2007, l'Union Européenne a coordonné une campagne de recensement de ces substances afin de les classer en fonction de leur dangerosité, via la mise en place du programme REACH. Cette directive vise à approfondir les connaissances scientifiques concernant les effets des substances chimiques sur la santé humaine et sur l'environnement, afin d'optimiser la gestion des risques liés à l'utilisation de ces produits. À terme, ce programme vise à l'interdiction et à la substitution dans l'Union européenne des substances chimiques les plus dangereuses, en particulier les substances CMR (cancérogènes, mutagènes, reprotoxiques), en prévoyant notamment des obligations à l'encontre des producteurs et importateurs de produits chimiques. Il revient donc aux industriels de démontrer l'innocuité des substances utilisées (Penman et al. 2015). Face à cette tâche, REACH incite notamment les scientifiques à développer et à mettre en place de nouvelles approches, dites *in silico*, complémentaires aux approches de toxicologie classiquement utilisées (*in vivo*, *ex vivo*, *in vitro*), afin d'améliorer et d'accélérer le criblage des substances potentiellement dangereuses.

1. Introduction

1.1. Le système endocrinien et ses hormones

Pour comprendre la problématique majeure soulevée par les perturbateurs endocriniens, il est au préalable nécessaire de connaître le fonctionnement et l'importance du système endocrinien chez l'Homme.

1.1.1. Qu'est-ce que le système endocrinien ?

Le système endocrinien (SE) est constitué d'organes et d'ensembles de cellules spécialisées dans l'élaboration de messagers chimiques, les hormones, qui régulent, à distance le plus souvent, un nombre important de processus physiologiques (Hiller-Sturmhöfel and Bartke 1998). Pour se faire, les organes endocriniens (également appelés glandes endocrines) ont pour rôle la production et le relargage d'hormones dans la circulation sanguine ou le système lymphatique. Les hormones ont ainsi la capacité d'agir spécifiquement sur des cibles (cellules, tissus ou organes) plus ou moins éloignées au sein de l'organisme. Certaines hormones peuvent parfois agir sur des cellules proches sans passage par la circulation générale (activité dite paracrine), voire même agir directement sur la cellule qui les a sécrétées (activité dite autocrine). Enfin certaines hormones peuvent également avoir une action intracrine en agissant au sein même de la cellule endocrine, sans avoir été relarguées dans le compartiment extracellulaire. Par ailleurs, certaines glandes du SE sont dites endo-exocrines, comme le pancréas ou les gonades. Il s'agit de glandes à activité mixte endocrine, par la sécrétion d'hormones dans la circulation sanguine, et exocrine, par l'expulsion de produits de sécrétion à l'extérieur de l'organisme via un canal excréteur. La Figure 1 liste de manière schématique les différents modes de communications des hormones.

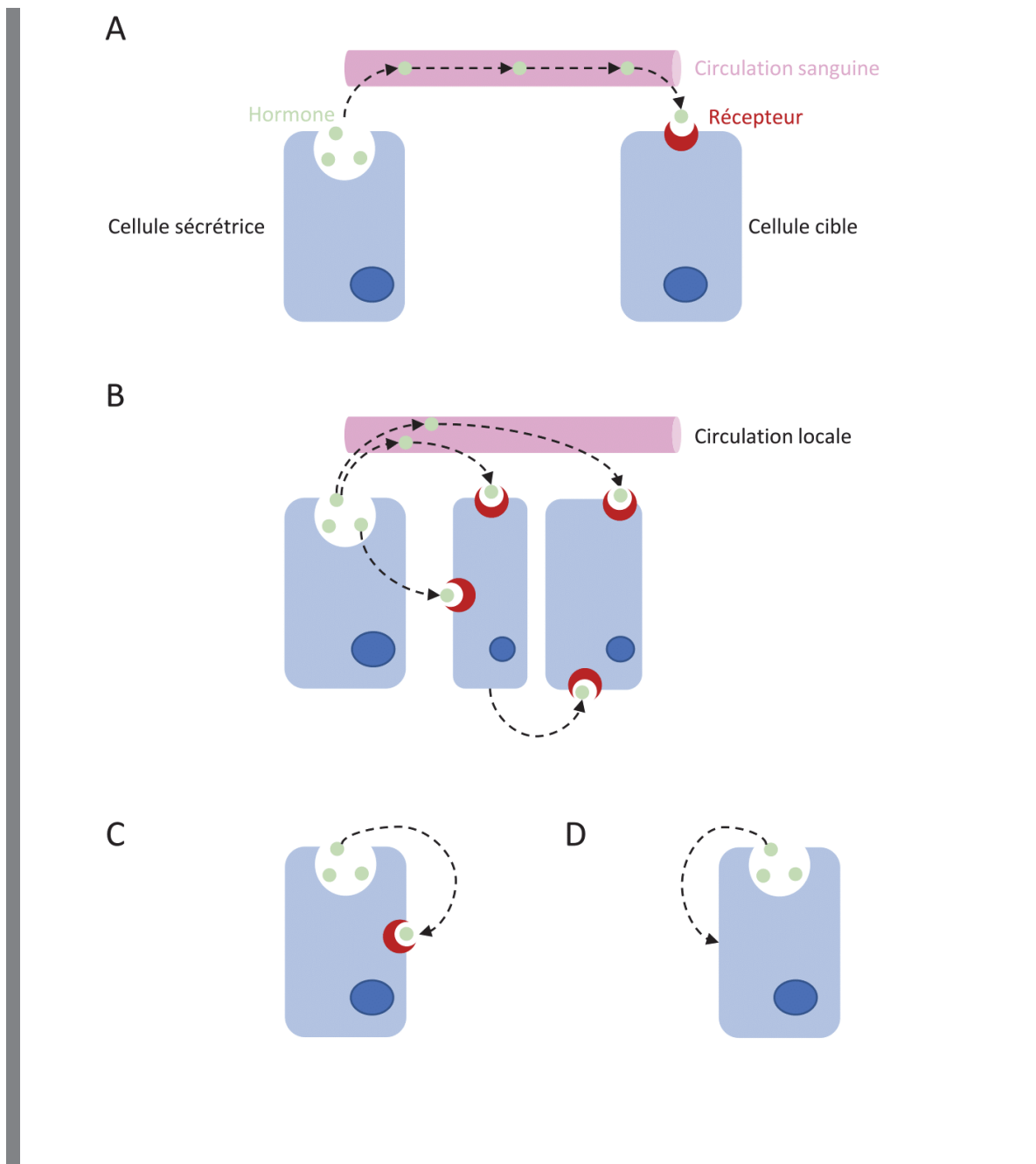


Figure 1. Modes de communications cellulaires.

Les hormones agissent de quatre manières différentes. A) De manière endocrine, l'hormone est relâchée dans la circulation sanguine afin d'atteindre la cellule cible. B) De manière paracrine, l'hormone agit sur des cellules cibles proches. C) De manière autocrine, l'hormone cible les récepteurs de la cellule sécrétrice. D) De manière intracrine, l'hormone agit sur la cellule sécrétrice sans passer par l'espace extracellulaire.

1.1.2. Présentation des différents types d'hormones

Par définition, une hormone est une molécule biologiquement active, sécrétée par une glande endocrine, régulant à distance et par voie sanguine des cellules cibles. Elles peuvent être classées en trois groupes selon leur nature biochimique, dont dépendra aussi leur mode d'action (Abiven, Raffin-Sanson, and Bertherat 2004).

1.1.2.1. Les hormones peptidiques

Les hormones peptidiques (Figure 2A) ont une taille allant de moins d'une dizaine à plusieurs centaines d'acides aminés. Elles sont, la plupart du temps, synthétisées dans le réticulum endoplasmique granuleux (REG) sous forme de préprohormones (PPH). Ces PPH sont soumises à l'activité enzymatique du REG et deviennent des prohormones, toujours inactives. Ces prohormones sont transportées vers l'appareil de Golgi par des vésicules de transports, puis elles subissent dans les granules sécrétoires (bourgeonnement de l'appareil de Golgi) une maturation enzymatique. Les prohormones sont alors maturées en hormones actives par des prohormones convertases. La vésicule formée par le granule sécrétoire permet à l'hormone de franchir la bicouche lipidique de la membrane plasmique. Une fois dans le sang, ces hormones ciblent les cellules d'intérêts via des récepteurs protéiques membranaires, assurant la transduction du signal à l'intérieur de la cellule.

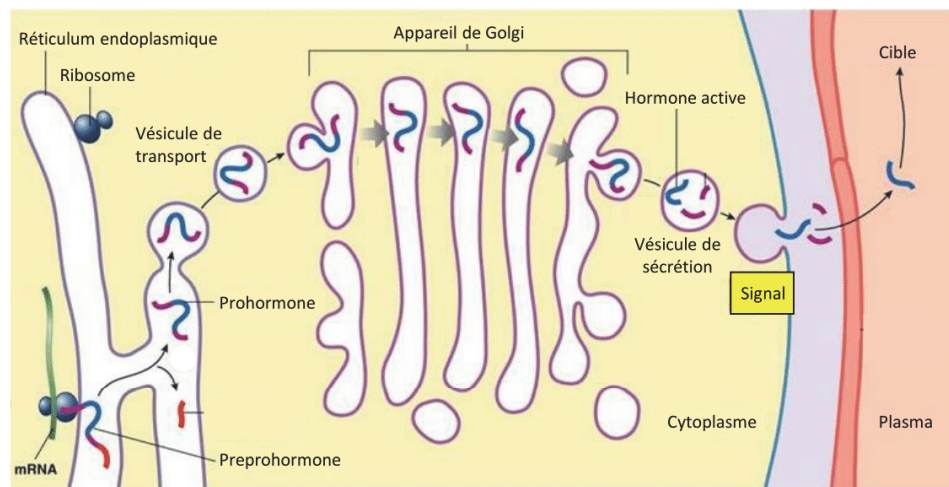
1.1.2.2. Les hormones stéroïdiennes

Les hormones stéroïdiennes sont des lipides dérivant du cholestérol et sont synthétisées dans le cytosol (mitochondrie et réticulum endoplasmique lisse) (Figure 2B). Leur caractère hydrophobe leur permet de traverser facilement la membrane plasmique. En revanche, ces hormones requièrent généralement de se complexer avec des protéines plasmatiques afin de circuler dans le sang. C'est au niveau des capillaires sanguins que l'hormone stéroïdienne est libérée de son transporteur pour pouvoir pénétrer ensuite dans le cytoplasme de la cellule cible et se lier aux récepteurs intracellulaires.

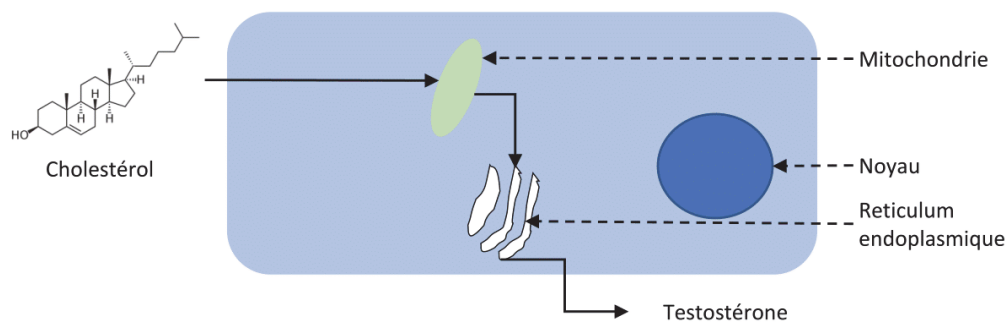
1.1.2.3. Les hormones monoamines

Les hormones monoamines sont des molécules de petite taille majoritairement dérivées de la tyrosine et du tryptophane incluant l'adrénaline, la noradrénaline, la dopamine et la mélatonine (Figure 2C). Comme les hormones peptidiques, ces hormones peuvent délivrer leurs signaux aux cellules cibles par l'intermédiaire de récepteurs transmembranaires spécifiques. D'autres hormones dérivées de la tyrosine comme la T3 et T4, agissent en revanche de la même manière que les hormones stéroïdiennes et se fixent à des récepteurs nucléaires.

A



B



C

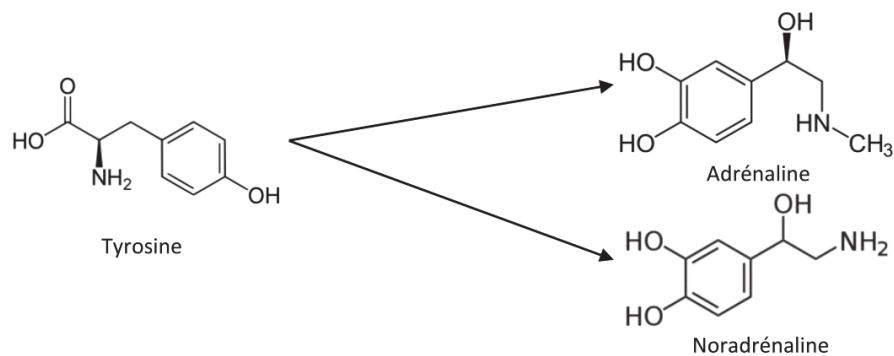


Figure 2. Schéma des différents types d'hormones

Il existe 3 grands types d'hormones. A) Les hormones peptidiques. Elles subissent une maturation enzymatique dans le réticulum endoplasmique et l'appareil de Golgi, passant du stade de préprohormones à prohormones pour terminer en hormones actives. B) Les hormones stéroïdiennes. Le précurseur de ces hormones est le cholestérol. Ce dernier subit un processus de conversion enzymatique avant de devenir une hormone. Ce processus est appelé stéroïdogénèse. C) les hormones monoamines. Ces hormones dérivent le plus souvent de la tyrosine.

1.1.3. Le mécanisme d'action des hormones

Afin que les différentes hormones puissent exercer leurs fonctions biologiques, il est nécessaire que des organes/tissus cibles soient capables d'intercepter et d'interpréter ces messages *via* des récepteurs cellulaires. Les hormones agissent ainsi de façon spécifique en se fixant de manière complémentaire à une cellule cible via des récepteurs hormonaux adéquats. Ces récepteurs sont présents soit au niveau de la membrane plasmique, soit à l'intérieur de la cellule (Figure 3A).

Les récepteurs hormonaux membranaires sont des protéines qui assurent la transduction du signal hormonal vers le cytoplasme, sans que l'hormone ne pénètre à l'intérieur de la cellule cible (Figure 3B). Ces récepteurs sont divisés en trois familles : les récepteurs couplés aux protéines G, les récepteurs tyrosine kinase et les récepteurs des cytokines. Après fixation de l'hormone à ces récepteurs deux types de réponses peuvent être générées : une réponse rapide et une réponse différée ou système de transduction. Le système de transduction intervient lorsque le récepteur est le début d'une chaîne moléculaire menant à la création d'un signal via des messagers secondaires. Ce système est traditionnellement composé de 3 éléments : un récepteur, un système de couplage protéique et une ou plusieurs protéines effectrices. Le système de couplage associant le récepteur à une ou plusieurs protéines effectrices est localisé sur la face interne de la membrane plasmique. Les protéines effectrices sont des enzymes qui, une fois stimulées, catalysent la synthèse de vecteurs intracellulaires porteurs de l'information amenée à la cellule cible par l'hormone. Ces vecteurs ou messagers secondaires tels que le calcium, l'AMPc (adénosine monophosphate cyclique), par exemple, provoquent de manière plus ou moins directe la phosphorylation de protéines présentes dans la cellule. Cette phosphorylation transforme leur activité biologique créant ainsi l'effet endocrinien recherché. Enfin, la réponse dite rapide de ces récepteurs hormonaux membranaires se fait grâce à l'ouverture de canaux ioniques présents dans la structure membranaire. L'ouverture de ces canaux crée un courant ionique, modifiant les propriétés électro-physiologiques des cellules cibles.

Les récepteurs hormonaux intracellulaires peuvent être cytosoliques ou bien nucléaires, et agissent directement en tant que facteur de transcription de certains gènes spécifiques (Bertherat 2004) (Figure 3C). On parle de récepteurs nucléaires dont l'activité est dépendante de la liaison de leur ligand. Les stéroïdes liposolubles se lient par exemple à des récepteurs cytosoliques, et c'est le complexe hormone/récepteur qui migre ensuite dans le noyau.

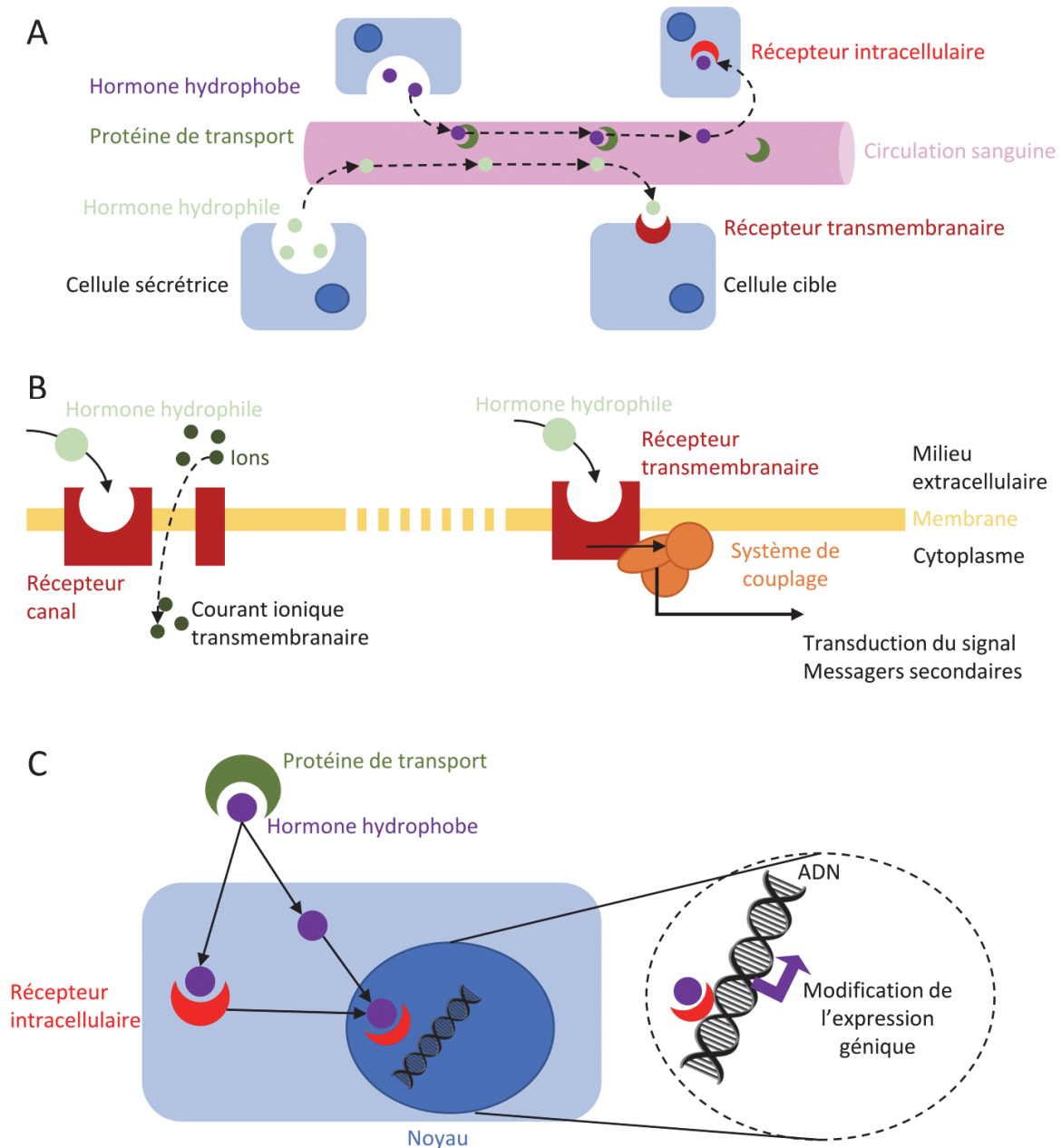


Figure 3. Mécanismes d'action des hormones

A) Représentation du mode d'action classique des hormones hydrophiles et lipophiles. B) Action des hormones sur les récepteurs membranaires. Activation des canaux ioniques : réponse rapide. Activation du système de couplage : action légèrement différée. C) Action des hormones sur les récepteurs intracellulaires.

À l'inverse, les hormones thyroïdiennes atteignent librement le noyau et reconnaissent des récepteurs nucléaires présents de manière constitutionnelle sur la séquence d'ADN cible. Une fois activé, le complexe hormone/récepteur agit sur la séquence d'ADN en se fixant sur la région promotrice de gènes cibles, modulant ainsi leur transcription. La réponse induite n'est pas immédiate, mais est durable dans le temps.

1.1.4. La régulation du système endocrinien

Le rôle principal du SE est de maintenir un état d'équilibre au sein de l'organisme, l'homéostasie, quel que soit l'environnement extérieur. Pour cela, la libération d'hormones dans le sang est modulée par trois principaux stimuli, nerveux, humoraux et hormonaux, chacun amenant les diverses glandes endocrines à produire et à libérer des hormones.

Stimuli nerveux : La fonction endocrinienne est étroitement régulée par le système nerveux par le biais de neurofibres qui stimulent la libération d'hormones. L'exemple type est celui de la réponse des glandes surrénales, directement contrôlées par le système nerveux, libérant de l'adrénaline et de la noradrénaline en réponse à un stress intense.

Stimuli humoraux : La libération des hormones peut également être régulée par une variation du taux de certains ions ou nutriments dans le sang, constituant le mécanisme de contrôle hormonal le plus simple. Ce système maintient par exemple en permanence la glycémie dans des taux physiologiquement acceptables. De la même façon, une carence en calcium dans le sang stimule la sécrétion de l'hormone parathyroïdienne, et, inversement, un taux élevé de calcium stimule la libération de calcitonine par la thyroïde.

Stimuli hormonaux : Enfin, la production et la libération de nombreuses hormones sont elles-mêmes régulées par d'autres hormones. On parle alors d'"axe" pour désigner l'ensemble des glandes et organes qui se régulent mutuellement. L'un de ces ensembles emblématiques est l'axe hypothalamo-hypophyso-gonadique, la GnRH sécrétée par l'hypothalamus stimulant la sécrétion de FSH et de LH par l'adénohypophyse. Ces deux hormones dites gonadotropes stimulent ensuite les gonades, qui exercent elles-mêmes un rétrocontrôle hormonal sur l'hypothalamus et l'adénohypophyse.

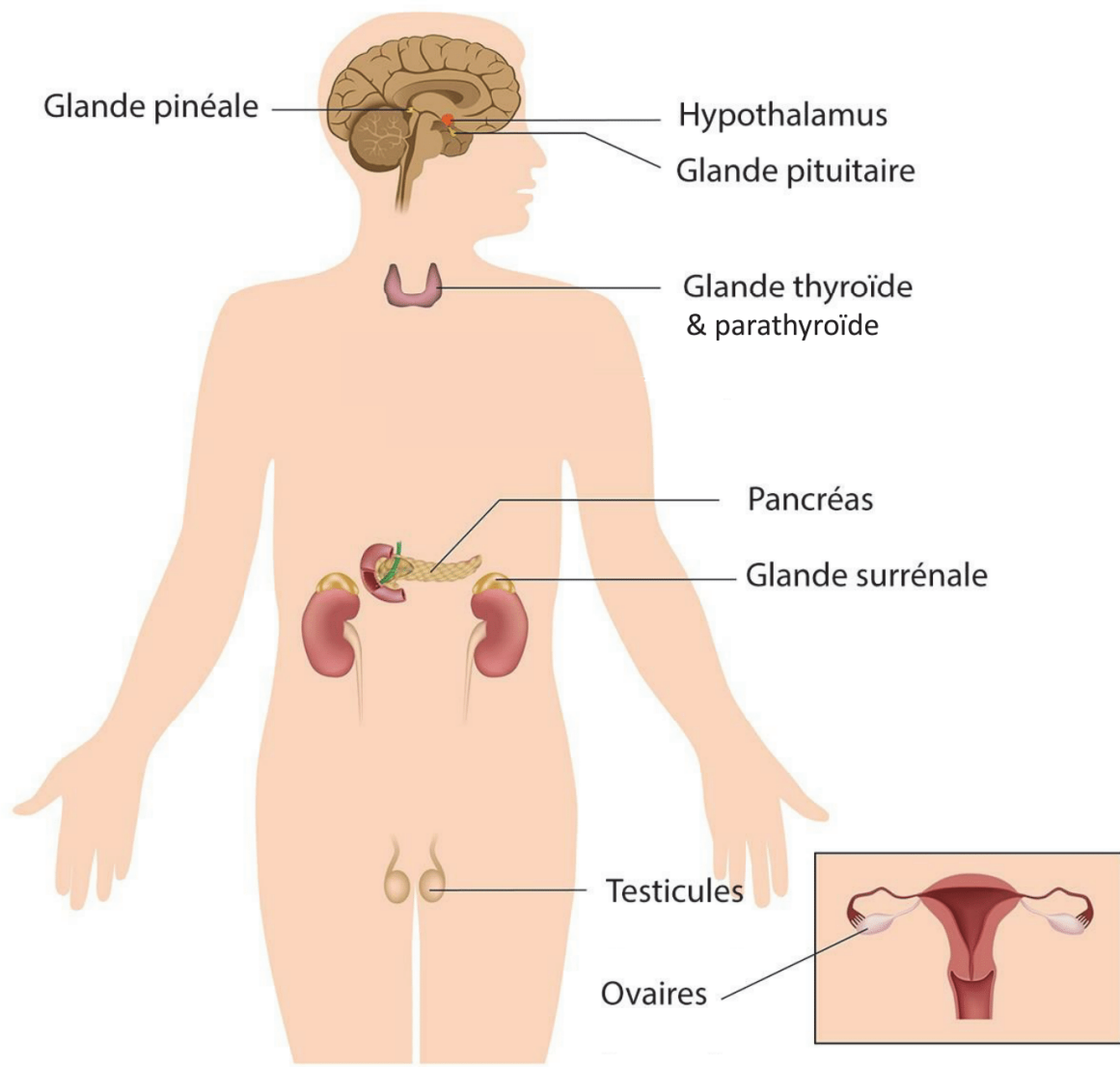


Figure 4. Schéma du système endocrinien

Représentation schématique du système endocrinien. Il est composé de 9 glandes : l'épiphyse ou glande pinéale, l'hypothalamus, l'hypophyse ou glande pituitaire, la thyroïde, les glandes parathyroïdiennes, le pancréas, les glandes surrénales et les gonades (ovaires et testicules)

1.1.5. Les glandes endocrines et leurs hormones

Les principales glandes endocrines sont l'épiphyse, l'hypothalamus, l'hypophyse, la thyroïde, les parathyroïdes, les surrénales, le pancréas et les gonades (Nussey et al. 2001). Dans une moindre mesure, l'estomac, l'intestin, les reins, le cœur et le placenta participent également à ce système (Figure 4).

1.1.5.1. L'épiphyse ou glande pinéale

L'épiphyse, ou glande pinéale est située au centre de l'encéphale, au-dessus et à l'arrière du troisième ventricule. Cette glande est constituée de cellules sécrétrices appelées pinéalocytes. Ces cellules sont impliquées dans la synthèse de la mélatonine, neurohormone sécrétée la nuit et ayant un rôle majeur dans le cycle circadien (Cassone et al. 1993). Cette hormone intervient dans différents processus biologiques dont la reproduction (rôle dans la synthèse et la fonction des stéroïdes comme les œstrogènes, la testostérone et la progestérone), la croissance cellulaire ou encore la régulation de la prise de poids via son action sur la glycémie, l'insulinémie ou encore les tissus adipeux stockés dans l'organisme (Reiter, Tan, and Fuentes-Broto 2010; Srinivasan et al. 2009; Prunet-Marcassus et al. 2003). De plus, par le biais de l'inhibition de l'absorption des acides gras par les cellules cancéreuses ou de l'activité de la télomérase, entraînant ainsi l'apoptose de cellules cancéreuses, la mélatonine est capable de bloquer l'angiogenèse et la prolifération tumorale (Mediavilla et al. 2010).

1.1.5.2. L'hypothalamus

L'hypothalamus est une structure du système nerveux central, située sur la face ventrale de l'encéphale. Cet organe est associé à l'hypophyse, formant ainsi le complexe hypothalamo-hypophysaire. Deux hormones sont produites par l'hypothalamus, stockées dans l'hypophyse puis relâchées dans le sang : l'hormone antidiurétique (ADH) ou vasopressine, et l'ocytocine. L'hypothalamus sécrète aussi des hormones dites libérines régulant les sécrétions hypophysaires :

- La **gonadolibérine** (GnRH ou LHRH) est sécrétée de façon pulsatile et agit sur les cellules gonadotropes de l'hypophyse. La fixation de la GnRH sur son récepteur induit une augmentation de la transcription de son propre gène et des gènes codants pour les hormones gonadotropes, c'est-à-dire la FSH et la LH (Counis et al. 2005).
- La **thyéolibérine** stimule la sécrétion de la thyroïdostimuline (TSH).

- La **somatocrinine** (GH-RH) permet la libération de l'hormone de croissance. La somatostatine (GH-IH) permet l'inhibition de l'hormone de croissance.
- La **corticolibérine** (CRF) stimule la sécrétion de la corticotrophine. Enfin, la dopamine inhibe la sécrétion de la prolactine.

1.1.5.3. L'hypophyse ou glande pituitaire

L'hypophyse est située à la base du cerveau et est sous le contrôle de l'hypothalamus, auquel elle est attachée. Elle assure la liaison entre le système nerveux et le SE. Son rôle physiologique est crucial pour l'organisme, puisqu'elle sécrète un grand nombre d'hormones contrôlant et régulant la fonction de nombreuses autres glandes endocrines. D'un point de vue anatomique, l'hypophyse est constituée de trois lobes : l'antéhypophyse (ou adénohypophyse) située en avant, le lobe intermédiaire et la posthypophyse (ou neurohypophyse) située en arrière (Figure 5).

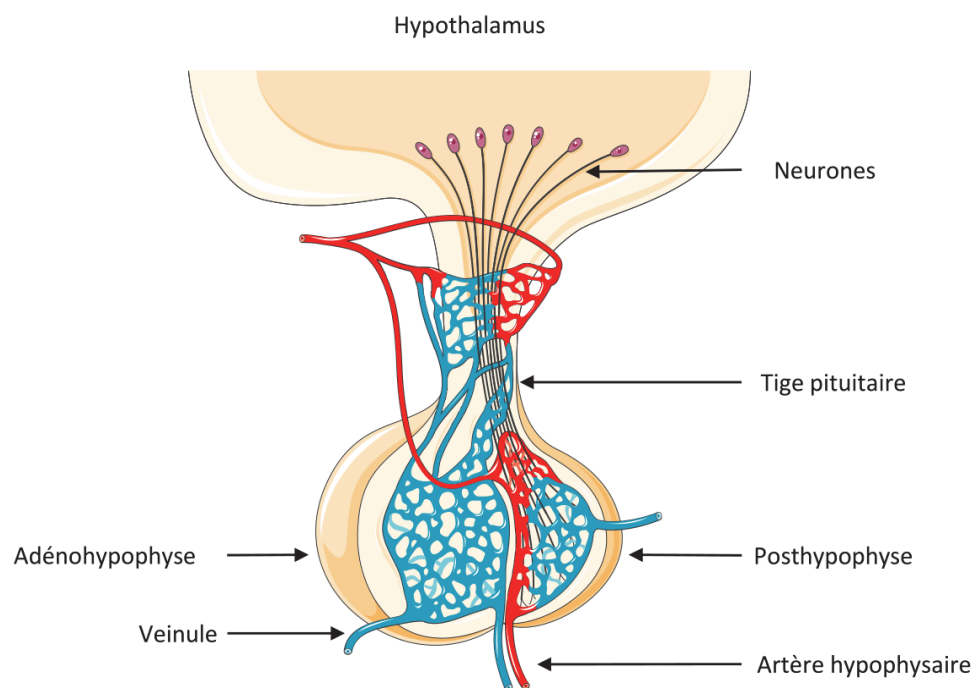


Figure 5. Schéma de l'hypothalamus et de l'hypophyse

Schéma de l'hypothalamus et l'hypophyse. Sont représentées ici les différentes parties de l'hypophyse, la tige pituitaire et les neurones.

- **L'antéhypophyse** est composée de nombreux types cellulaires qui produisent et libèrent de nombreuses hormones adénohypophysaires, ayant chacune des effets physiologiques distincts sur l'organisme. 1) **L'hormone adrenocorticotrope (ACTH)** qui stimule le cortex des glandes surrénales, induisant la synthèse de corticostéroïdes tels que les glucocorticoïdes et les androgènes à partir du cholestérol. 2) **La thyroestimuline (TSH)** stimule le développement et l'activité de sécrétion de la thyroïde. 3) **L'hormone folliculo-stimulante (FSH)** et **l'hormone lutéinisante (LH)** stimulent les glandes sexuelles. Ces hormones gonadotropes (ou gonadotrophines) sont produites par les cellules gonadotropes et régissent le fonctionnement des gonades. Ces deux hormones sont des hétérodimères constitués d'une sous-unité commune, la glycoprotéine α (gp α), et d'une sous-unité β spécifique de chacune des deux hormones (FSH β , LH β) (Jiang, Dias, and He 2014). Chez l'homme et la femme, la FSH stimule la production de gamètes (spermatozoïdes et ovules, respectivement), tandis que la LH favorise la production des hormones gonadiques. L'hormone lutéotrope est sécrétée par les cellules lactotropes et participe à la croissance des glandes mammaires. 4) **L'hormone de croissance (GH)** est l'hormone protéique la plus abondante sécrétée par l'adénohypophyse. Bien que la GH provoque la croissance et la division de la plupart des cellules de l'organisme, ses principales cibles sont les os et les muscles squelettiques. Elle régule le métabolisme de nombreux organes, dont le foie, l'intestin et le pancréas, et agit également sur les fonctions gonadiques comme la spermiogenèse, l'accroissement de la mobilité des spermatozoïdes et la stimulation de la synthèse des androgènes (Hull and Harvey 2000).

- **Le lobe intermédiaire** sécrète **l'hormone mélanotrope (MSH)** qui, en induisant la synthèse de mélanine, régule l'intensité de la pigmentation de la peau induite par les UVA.

- **La posthypophyse** stocke les hormones produites par l'hypothalamus et les distribue sous forme de neurohormones. Parmi celles-ci, la **vasopressine**, ou hormone **antidiurétique (ADH)**, régule le débit sanguin rénal et la filtration glomérulaire rénale, modulant ainsi la pression artérielle. L'ocytocine quant à elle, stimule les contractions utérines et celles des canaux galactophores lors de l'accouchement et de l'allaitement.

1.1.5.4. La thyroïde

La glande thyroïde est située dans la partie antérieure du cou et repose sur la trachée. Les hormones thyroïdiennes (**thyroxine (T4)**, et **triiodothyronine (T3)**) stimulent le métabolisme, régulent la croissance et la maturation des tissus de l'organisme et peuvent affecter la vigilance et l'humeur. La thyroïde sécrète également la **calcitonine**, qui abaisse le taux sanguin de calcium (calcémie) et est un antagoniste direct de la parathormone sécrétée par les glandes parathyroïdes (de Paula and Rosen 2010).

1.1.5.5. Les glandes parathyroïdes

Les glandes parathyroïdes sont de petites glandes rattachées à la face postérieure de la glande thyroïde. Elles sont généralement au nombre de quatre. Ces glandes sécrètent la parathormone ou **hormone parathyroïdienne (PTH)**. Cette hormone peptidique joue un rôle majeur dans la régulation osseuse, rénale ou duodénale du calcium (Talmage and Mobley 2008). La PTH est déversée quand la concentration sérique de calcium diminue. À l'inverse, l'hypercalcémie inhibe sa production.

1.1.5.6. Les glandes surrénales

Les glandes surrénales sont des glandes triangulaires situées au-dessus du rein, dans la cavité péritonéale. Ces glandes sont divisées en deux structures anatomiques : la corticosurrénale, périphérique, constituant 80 à 90% de la glande surrénale, et la médullosurrénale, centrale (Rosol et al. 2001).

La corticosurrénale est divisée en trois couches distinctes qui sécrètent trois groupes d'hormones différents. **1)** Les **minéralocorticoïdes** sont en majorité représentés par l'aldostérone qui est chargée de réguler l'équilibre des ions sodium. L'aldostérone, via un échangeur sodium/potassium, stimule la réabsorption des ions sodium dans l'urine afin de diminuer les concentrations urinaires et plasmatiques de potassium, et moduler ainsi la pression sanguine (Catena et al. 2014). **2)** Les **gonadocorticoïdes** sont composés principalement d'androgènes (Labrie 2004). Enfin, **3)** les **glucocorticoïdes** ont un rôle crucial dans la réponse au stress cellulaire et l'adaptation de l'organisme aux variations environnementales. Parmi ces glucocorticoïdes, le cortisol, hyperglycémiant, favorise la synthèse de glucose dans le foie et la production d'anticorps.

La médullosurrénale produit deux catécholamines : l'**adrénaline** et la **noradrénaline**. Ces hormones ont une action stimulante sur le cœur, augmentent la pression sanguine en agissant sur la constriction des vaisseaux sanguins, et augmentent la contractilité musculaire, permettant ainsi à l'organisme de réagir de manière adaptée en situation de stress (Kvetnansky, Sabban, and Palkovits 2009).

1.1.5.7. Le pancréas

Le pancréas est situé à l'arrière de l'estomac et assure une fonction exocrine par la libération des sucs pancréatiques au sein du duodénum. Répartis tout du long de cet organe, les îlots de Langerhans assurent eux sa fonction endocrine et sont composés de deux grandes populations de cellules : les cellules α synthétisent le **glucagon**, provoquant la libération du glucose dans le foie ; et les cellules β , plus nombreuses, produisent l'**insuline**, hormone agissant sur le métabolisme glucidique, protidique et

lipidique. Outre son rôle hypoglycémiant, l'insuline favorise en effet également la formation de protéines et le stockage lipidique.

1.1.5.8. Les gonades

Les ovaires sont les glandes génitales femelles. Au même titre que les testicules sont le siège de la spermatogenèse, ils assurent une fonction exocrine, l'ovogenèse. Par ailleurs, de façon cyclique, l'ovaire produit des hormones sexuelles : les œstrogènes (œstrone, œstradiol, l'estriol) et la progestérone sous l'influence de la FSH et de la LH sécrétées par l'hypophyse, assurant ainsi la régulation du système reproducteur (Richards and Pangas 2010). Les hormones progestatives sont produites par le corps jaune ovarien. Elles agissent sur l'activité sécrétoire de l'endomètre et jouent un rôle dans le développement de la glande mammaire. L'œstrogène sécrété est nécessaire pour le développement des organes reproducteurs, mais également pour l'acquisition des caractères sexuels secondaires comme l'ouverture du bassin, le développement de la poitrine et la pilosité pubienne et axillaire. Enfin, les ovaires produisent de l'inhibine, hormone régulatrice inhibant la sécrétion de la FSH, mais également de l'activine, hormone stimulant la production de FSH, et la relaxine, hormone provoquant l'assouplissement et la relaxation de l'utérus lors de l'accouchement.

Enfin, le testicule, glande génitale mâle, est volontairement omis ici, car détaillé plus bas (cf. section 1.2.).

1.1.5.9. Autres organes

D'autres tissus dans l'organisme assurent une fonction endocrine. Parmi ces tissus, le placenta joue un rôle important lors de la grossesse en produisant un certain nombre d'hormones stéroïdiennes. Il prend aussi en charge une partie des fonctions endocriniennes hypophysaires en sécrétant la gonadotrophine chorionique (hCG) (Acevedo 2002). Le tractus gastro-intestinal a également une fonction endocrine en sécrétant de la gastrine au niveau de l'estomac et de la cholécystokinine au niveau de l'intestin. Les reins sécrètent la rénine et l'érythropoïétine qui augmentent respectivement la pression sanguine et la synthèse des globules rouges par la moelle osseuse (Peart 1977). Le cœur assure une fonction endocrine par la sécrétion de facteur natriurétique auriculaire, agissant dans la régulation de la pression sanguine et de l'équilibre du sel et de l'eau dans l'organisme (Sagnella 2002). Enfin, les os par la production d'ostéocalcine se fixant sur les récepteurs des cellules de Leydig, régulent l'expression d'enzymes nécessaires à la synthèse de testostérone (3 β -HSD, Cyp17, StAR) et jouent un rôle sur la production de testostérone chez l'homme (Oury et al. 2011).

Système endocrinien	Hormones secrétées	Abréviations	Effets
Axe Hypothalamo-Hypophysaire	Hormone thyroïdienne	TRH ou PRH	Stimule la sécrétion de TSH et PRL au niveau de l'hypophyse
	Dopamine	DA	Inhibe la sécrétion de PRL au niveau de l'hypophyse
	Hormone de libération de l'hormone de croissance	GHRH	Stimule la sécrétion de GH au niveau de l'hypophyse
	Gonadolibérine	GnRH ou LHRH	Stimule la sécrétion de FSH et LH au niveau de l'hypophyse
	Corticolibérine	CRH	Stimule la sécrétion d'ACTH au niveau de l'hypophyse
	Hormone de croissance	GH	Stimule la croissance et la reproduction cellulaire
	Prolactine	PRL	Stimule la lactation
	Hormone folliculo-stimulante	FSH	Stimule la maturation des follicules chez la femelle et régule la spermatogénèse chez le mâle
	Hormone lutéinisante	LH	Stimule la sécrétion d'œstrogènes et induit l'ovulation chez la femelle. Stimule la sécrétion de testostérone chez le mâle
	Thyréostimuline	TSH	Stimule la sécrétion d'hormones thyroïdiennes
	Hormone adréno-corticotrope	ACTH	Stimule la sécrétion des glucocorticoïdes
	Mélano-stimuline	MSH	Stimule la production de mélanine
	Ocytocine		Stimule la contraction utérine et la lactation
	Vasopressine	ADH	Inhibe la sécrétion d'urine et maintient l'équilibre électrolytique
	Cortisol		Régule le métabolisme, la croissance et la maturation des tissus de l'organisme
	Déhydroépiandrostérone	DHEA	Action contre le vieillissement
	Adrénaline		Accélère le rythme cardiaque suite à un stress ou une activité physique
	Noradrénaline		Action dans l'attention, le sommeil et l'apprentissage
	Triiodothyronine	T3	Stimule la production de l'ARN polymérase I et II, la fréquence cardiaque et la force de contraction, la production de myéline, les neurotransmetteurs et la croissance axonale
Thyroïdes	Thyroxine	T4	Régule la vitesse du métabolisme et des processus de croissance et de différenciation des tissus
	Calcitonine		Régule le métabolisme du phosphore et du calcium
	Parathormone	PTH	Régule le métabolisme phospho-calcique du sang
	Hormone antimüllérienne	AMH	Intervient dans la régression des canaux de Müller durant l'embryogénèse
Gonades	Inhibine		Inhibe la synthèse de la FSH hypophysaire
	Activine		Active la synthèse de la FSH hypophysaire
	Insuline like 3	INSL3	Impliqué dans la descente testiculaire et le maintien de la spermatogénèse
	Œstrogènes	E1, E2 ou E3	Stimule le développement des organes reproducteurs et des caractères secondaires féminins, régule le cycle ovarien
	Progestérone	P4	Prépare l'endomètre à l'implantation de l'œuf fécondé, maintient le col de l'utérus fermé, intervient dans le développement des glandes mammaires
	Androstènedione		Participe à la synthèse des œstrogènes
	Inhibine		Inhibe la synthèse de la FSH hypophysaire
	Glucagon		Régule la glycémie
	Insuline		Régule les substrats énergétiques comme le glucose, les acides gras et les corps cétoniques
	Somatostatine		Inhibe la sécrétion de GH, de TSH, d'insuline et du glucagon
Autres glandes endocrines	Hormone chorionique gonadotrope	hCG	Maintien du corps jaune durant la grossesse et de la sécrétion de progestérone
	Estriol	E3	Forme principale d'œstrogènes produite pendant la grossesse
	Progestérone	P4	Maintien de la grossesse
	Mélatonine		Agit sur la reproduction, le système immunitaire et en tant qu'antioxydant
	Thymuline		Stimule l'immuno-compétence des lymphocytes T
	Thymopoïétine		Agit sur les cellules nourricières des prothymocytes

Tableau I. Tableau récapitulatif des différentes hormones et leurs actions

Hormones peptidiques - Hormones stéroïdes - Hormones monoaminées

Les principaux organes et glandes du SE, les hormones qu'ils sécrètent, et rôles physiologiques qu'elles assurent sont résumés dans le Tableau I. Ce tableau permet de constater rapidement de la complexité du SE et de son impact sur l'organisme tout au long de la vie.

1.2. Le testicule

Comme les ovaires, le testicule fait partie des glandes du SE. Durant ma thèse, le testicule a été l'organe sur lequel je me suis particulièrement focalisé. Voilà pourquoi je décrirais ici plus en détail le testicule, sa fonction endocrine et la régulation de cette fonction.

1.2.1. Généralités sur l'anatomie du testicule adulte

Le testicule est garant de la fonction de reproduction chez l'homme, fonction dépendante de sa capacité à produire des gamètes en nombre suffisant. Mais le testicule doit également assurer le bon développement du tractus génital mâle durant les premières années, et par la suite assurer le maintien de sa fonctionnalité via la sécrétion d'un niveau constant d'androgènes. Le testicule est composé de deux compartiments histologiques distincts : les tubes séminifères, qui représentent 90 % du volume testiculaire, où a lieu la spermatogenèse (fonction exocrine), et l'interstitium (ou tissu interstitiel) entre les tubes séminifères, siège de la stéroïdogénèse (fonction endocrine). Le testicule est logé dans une enveloppe de peau appelée scrotum, et entouré d'une capsule épaisse, riche en tissu conjonctif et fibres de collagènes, appelée albuginée. La partie supérieure de cette capsule s'épaissit et s'enfonce à l'intérieur du testicule pour former le corps d'Highmore. De là partent des cloisons conjonctives radiales, les septa testis, délimitant 200 à 300 lobules testiculaires plus ou moins triangulaires. Chaque lobule contient 2 à 4 tubes séminifères, où se divisent et se différencient les cellules germinales jusqu'à la formation des spermatozoïdes. Les espaces entre ces tubes sont occupés par un tissu interstitiel conjonctif lâche, riche en vaisseaux sanguins et lymphatiques et en fibres nerveuses, contenant notamment des cellules glandulaires de type endocrine, les cellules de Leydig. Les tubes séminifères se prolongent par de petits tubes courts et rectilignes appelés tubes droits, qui pénètrent ensuite dans le corps de Highmore et forment un réseau de canalicules, le *rete testis*. Ce dernier se poursuit par des tubes efférents établissant le lien entre le *rete testis* et l'épididyme qui coiffe le testicule. Le testicule est suspendu dans le sac scrotal par le cordon spermatique qui contient le canal défèrent, des vaisseaux sanguins et lymphatiques, et des fibres nerveuses. La Figure 6 décrit l'anatomie du testicule humain.

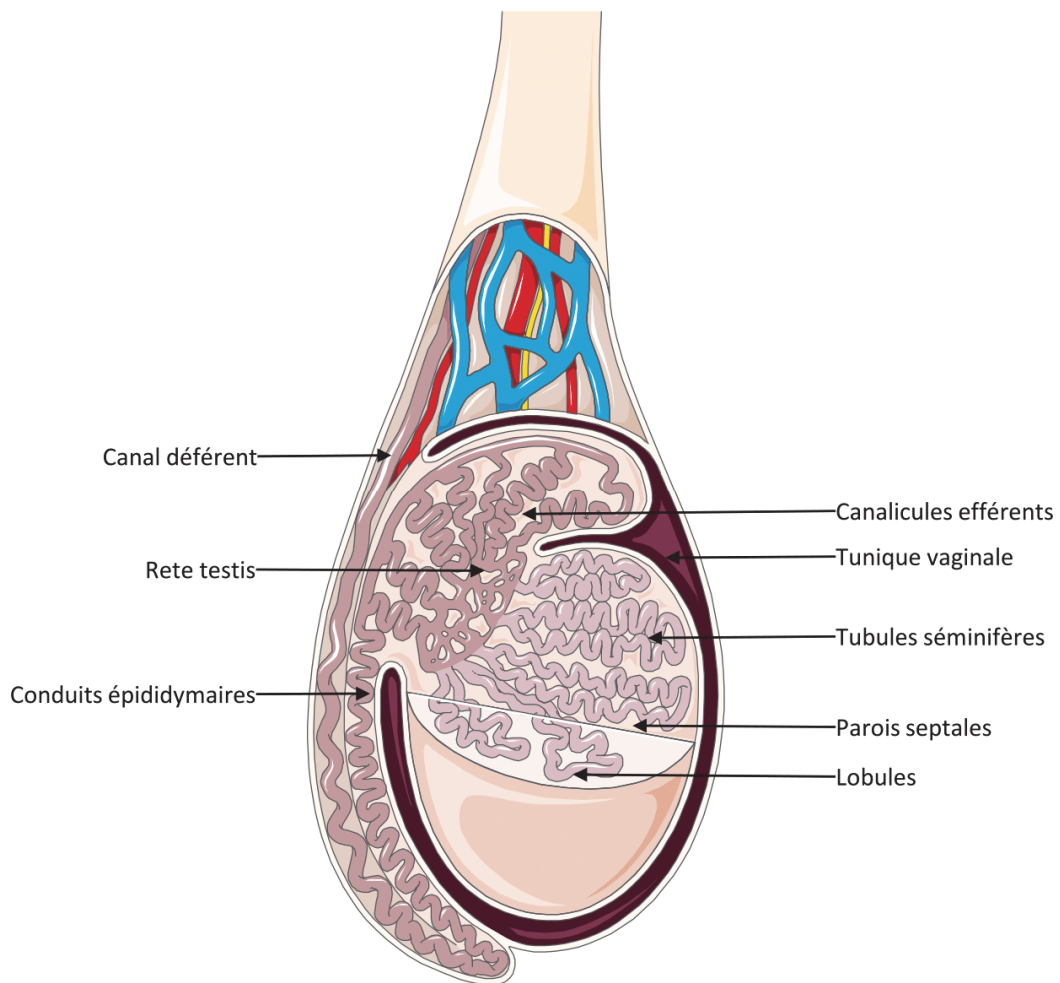


Figure 6. Anatomie du testicule adulte humain

Schéma représentatif du testicule humain, présentant : le canal déférent, le rete testis, les conduits épидидymaires, les canalicules efférents, la tunique vaginale, les tubules séminifères, les parois septales et les lobules.

1.2.2. Fonction exocrine du testicule

L'essentiel de mon travail ayant eu pour objet la fonction endocrine du testicule, je ne décrirai que brièvement sa fonction exocrine. La spermatogenèse est la fonction exocrine du testicule, et correspond à l'ensemble des transformations aboutissant à la formation d'un spermatozoïde mature. Ce phénomène prend place au sein des tubules séminifères contenant trois types de cellules : les cellules péricubulaires, les cellules de Sertoli et les cellules de la lignée germinale (Figure 7).

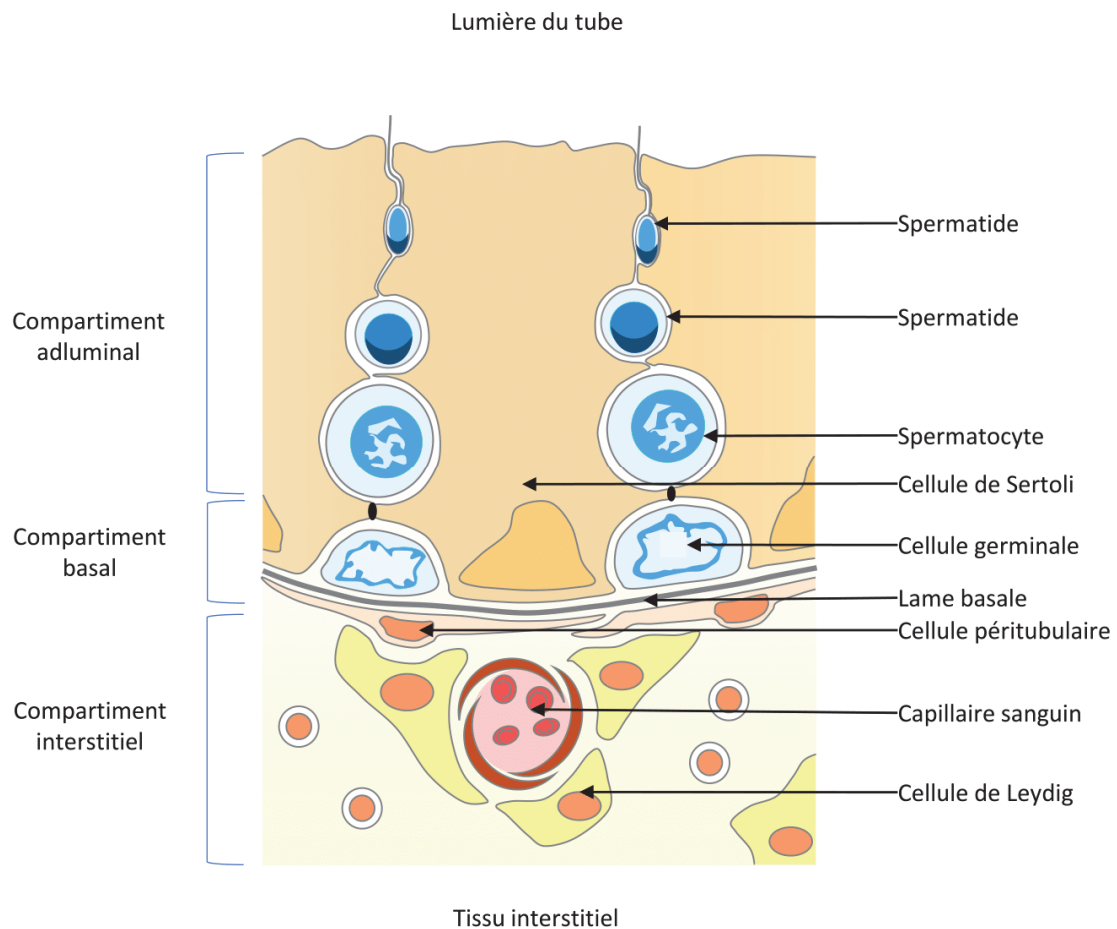


Figure 7. Coupe d'un tubule séminifère

Représentation schématique des tubes séminifères et du tissu interstitiel et des différents types cellulaires composant ces deux compartiments.

1.2.2.1. Les cellules péricubulaires

Les cellules péricubulaires sont des cellules allongées, possédant des caractéristiques proches des cellules musculaires lisses et/ou de fibroblastes, plus généralement dénommées myofibroblastes. Elles constituent un des éléments de la barrière hémato-testiculaire puisant dans la circulation sanguine le rétinol (vitamine A) qui sera redistribué aux cellules de Sertoli afin d'assurer le mécanisme de spermatogénèse (J. T. Davis and Ong 1995). Les caractéristiques partagées par les cellules péricubulaires avec les myofibroblastes, leur permettent de jouer un rôle dans la contraction des tubules séminifères. Cette contraction permet le transport des spermatozoïdes hors du tubule séminifère. En contact avec la surface basale des cellules de Sertoli, il a été démontré que les cellules péricubulaires interagissaient avec

les cellules de Sertoli pour produire une matrice extracellulaire nécessaire à l'intégrité structurale des tubules séminifères (E. W. Thompson, Blackshaw, and Raychoudhury 1995). Enfin les cellules péritubulaires jouent également un rôle dans la régulation des cellules souches spermatogoniales (Oatley and Brinster 2012) et pourraient être impliquées dans le développement postnatal des testicules (Nurmio et al. 2012).

1.2.2.2. Les cellules de Sertoli

Les cellules de Sertoli sont les cellules nourricières de la spermatogénèse. Elles apportent un support structural et nutritionnel pour les cellules germinales en développement. Elles partent de la membrane basale et s'étendent jusqu'à la lumière du tube. Les cellules de Sertoli sont caractérisées par une structure triangulaire et entourent les cellules germinales. Ces cellules remplissent différents rôles (Mruk and Cheng 2004). Tout d'abord, les cellules de Sertoli apportent un support physique pour le maintien des cellules germinales. Elles jouent également un rôle important dans la phagocytose et l'élimination des cellules germinales dégénérantes et des corps résiduels issus des spermatides durant la spermatogénèse. Les cellules de Sertoli forment entre elles des jonctions serrées établissant un élément crucial de la barrière hémato-testiculaire séparant l'épithélium séminifère en un compartiment basal et un compartiment adluminal. Ces cellules sont également impliquées dans le mouvement des cellules germinales ainsi que de la libération des spermatides matures dans la lumière du tube séminifère (spermiation). Enfin, les cellules de Sertoli sécrètent de nombreuses substances. Parmi celles-ci, l'activine et l'inhibine, hormones régulant positivement ou négativement la sécrétion de FSH, respectivement. Mais elles sont également la cible d'hormones telles que la testostérone. Il a été démontré que la régulation des cellules de Sertoli par la testostérone est capitale pour le processus de spermatogénèse (Walker 2011). Par ailleurs il est important de noter que la production de spermatozoïdes pour une espèce donnée étant directement corrélée au nombre final de cellules de Sertoli, le contrôle hormonal via la FSH et l'insuline (Pitetti et al. 2013) de la prolifération des cellules de Sertoli dans la gonade en développement est primordiale pour la future production des cellules germinales (Petersen and Soder 2006).

1.2.2.3. Les cellules de la lignée germinale

À partir de la puberté et tout au long de la vie, les cellules de la lignée germinale mâle se différencient en spermatozoïdes selon un processus de différenciation cyclique appelé spermatogénèse. Ce processus a une durée d'environ 74 jours chez l'homme (Figure 8) et implique trois phases :

- Phase mitotique : Les cellules germinales effectuant la mitose sont appelées spermatogonies. On distingue tout d'abord les spermatogonies de type A dites « dark », en contact avec la membrane basale, et qui représentent les cellules souches de la spermatogénèse. Elles se divisent de manière asymétrique pour donner à nouveau une spermatogonie de type A « dark », assurant ainsi le renouvellement du stock de cellules souches, et une spermatogonie de type A « pale », en charge d'initier la spermatogénèse. Chaque spermatogonie de type A « pale » va en effet se diviser pour donner deux spermatogonies de type B, lesquelles se divisent à leur tour et donnent deux spermatocytes qui sont dès lors engagés en méiose (Clermont 1963).

- Phase méiotique : Les spermatocytes subissent la méiose : ils vont donc enchaîner deux divisions cellulaires pour une seule synthèse d'ADN et former ainsi quatre nouvelles cellules haploïdes appelées spermatides. Outre la réduction du matériel génétique, la méiose implique également un brassage de l'information par le biais de recombinaisons homologues entre les chromosomes d'origines maternelles et paternelles.

-Phase post-méiotique ou spermiogénèse : La spermiogénèse ne fait pas intervenir de division cellulaire, mais correspond à une transformation morphologique profonde au cours de laquelle les spermatides se différencient en une entité hautement différenciée et douée de mobilité, le spermatozoïde. Cela nécessite notamment la formation de l'acrosome (à partir de l'appareil de Golgi), une compaction nucléaire extrême, le développement d'un flagelle, et l'élimination de la majeure partie du cytoplasme et des organites intracellulaires.

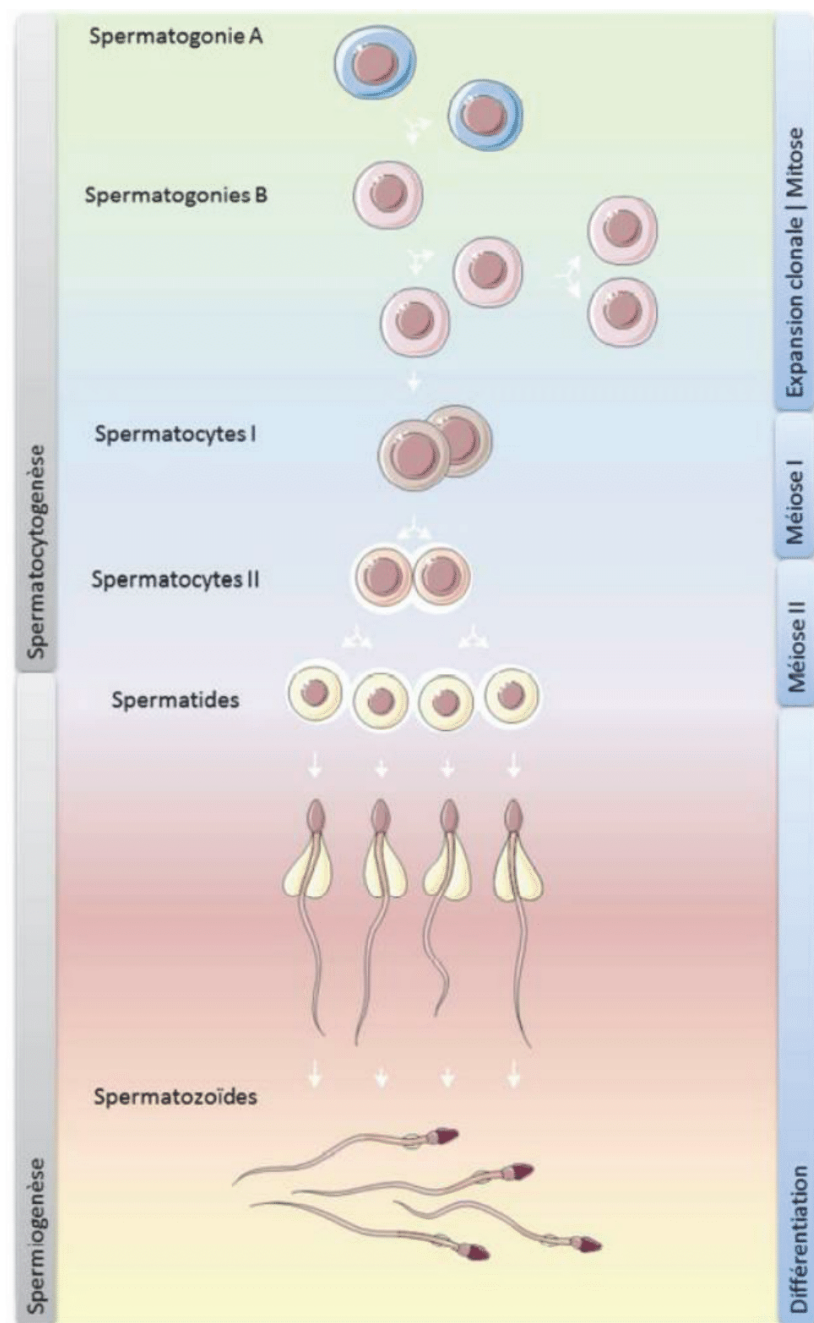


Figure 8. Fonction exocrine du testicule : la spermatogénèse

D'après (B Jégou 1995). La spermatogénèse correspond au phénomène de différenciation des cellules germinales en spermatozoïdes. Les spermatogonies de type A se divisent par mitose : une des cellules filles renouvelle le stock de spermatogonies de type A, l'autre devient une spermatogonie du type B. Celles-ci vont devenir des spermatocytes et subir la méiose. Après deux divisions méiotiques, les spermatides formées se différencient en spermatozoïdes.

1.2.3. Fonction endocrine du testicule

1.2.3.1. Le tissu interstitiel

Cet espace est composé de vaisseaux sanguins et lymphatiques, de cellules de Leydig, de fibroblastes et de macrophages. Parmi ces différents types cellulaires, les cellules de Leydig, qui synthétisent à la fois de la testostérone et l'hormone peptidique Insulin-like factor 3 (l'INSL3), jouent un rôle capital pour la différenciation gonadique et le maintien de la fonction reproductive.

Le niveau de testostérone étant directement corrélé à l'activité des cellules de Leydig, l'étude de ce taux a permis mettre en évidence la présence d'une maturation des cellules de Leydig en trois étapes. La première intervient pendant la vie fœtale. La seconde intervient 2 à 3 mois après la naissance. Cette seconde étape est la conséquence directe de la réactivation de l'axe hypothalamo-hypophyso-gonadique provoquant une élévation des taux circulants de LH associée à une augmentation des niveaux de testostérone. Enfin, la troisième vague de maturation a lieu durant la puberté. Pendant toute cette période, des cellules de Leydig immatures, mais exprimant la machinerie stéroïdogénique, se développent dans le tissu interstitiel. Toutes ces étapes de différenciation impliquent un grand nombre de phénomènes biochimiques et de changements morphologiques importants conduisant à l'expression des enzymes de la stéroïdogenèse et du récepteur à la LH (LHR).

L'INSL3 est connue pour son rôle essentiel dans la descente testiculaire au cours de la vie fœtale (Parada and Nef 1999). Plus tard dans la vie, l'augmentation du niveau sérique d'INSL3 est utilisée comme marqueur d'apparition de la puberté chez les jeunes garçons. Le niveau de cette hormone restera relativement élevé chez adulte. Produit exclusivement par les cellules de Leydig, l'INSL3 est une des hormones peptidiques circulantes majeures chez l'homme adulte. Sa production diffère de la régulation dite classique par les facteurs hormonaux, et donc de l'axe hypothalamo-hypophyso-gonadique. En revanche, toute affection de la condition des cellules de Leydig aura des conséquences sur les niveaux circulants d'INSL3 (Ivell and Anand-Ivell 2009). On attribue à l'INSL3 un rôle dans le maintien de la spermatogenèse, du fait de la présence de son récepteur RXFP2 sur les cellules germinales pré et post-méiotiques. Enfin plusieurs études tendent également à prouver que l'INSL3 pourrait jouer un rôle dans la prévention de l'apoptose des cellules germinales (Kawamura et al. 2004) et dans le métabolisme osseux, notamment dans la réduction de l'incidence de l'ostéoporose chez l'homme (Ferlin et al. 2008).

1.2.3.2. La stéroïdogénèse

La stéroïdogénèse, mécanisme de synthétisation des stéroïdes, est essentielle au bon déroulement de la fonction de reproduction. Dans le testicule, ce processus a lieu au niveau des cellules de Leydig. Cette biosynthèse nécessite l'intervention de plusieurs enzymes agissant en cascade à partir d'un précurseur commun : le cholestérol. Ce dernier pénètre dans la mitochondrie via des transporteurs StAR (Steroidogenic Acute Regulatory protein) » (Stocco 2000; Robertson et al. 2016) où se déroulera le processus de transformation du cholestérol en testostérone. La première étape est la conversion du cholestérol en prégnénolone, grâce à une enzyme de la famille des cytochromes P450, le cytochrome P450_{scc} (P450 *cholestérol side chain clivage*) ou Cyp11A1, qui coupe la chaîne latérale de la molécule de cholestérol. Puis, la prégnénolone est transportée vers le réticulum endoplasmique lisse où elle est transformée en progestérone par la 3 β -HSD (3 beta hydroxysteroid deshydrogenase). Cette enzyme est également impliquée dans la conversion des Δ^5 -3 β -hydroxystéroïdes en Δ^4 -3-kétostéroïdes, étape essentielle dans la biosynthèse des hormones stéroïdiennes biologiquement actives. La progestérone est ensuite convertie en 17 α -OH-progestérone puis en androstènedione par le cytochrome P450_{c17} (ou Cyp17A1). Enfin, la 17 β -HSD (17 bêta hydroxysteroid deshydrogenase) transforme l'androstènedione en testostérone (Figure 9).

Si la testostérone (de même que l'androstènedione) peut être métabolisée sous l'action de l'aromatase en œstradiol (Conley and Hinshelwood 2001), elle peut également être métabolisée de façon irréversible par la 5 α -réductase en un dérivé plus actif, la dihydrotestostérone (DHT) (Mahendroo and Russell 1999).

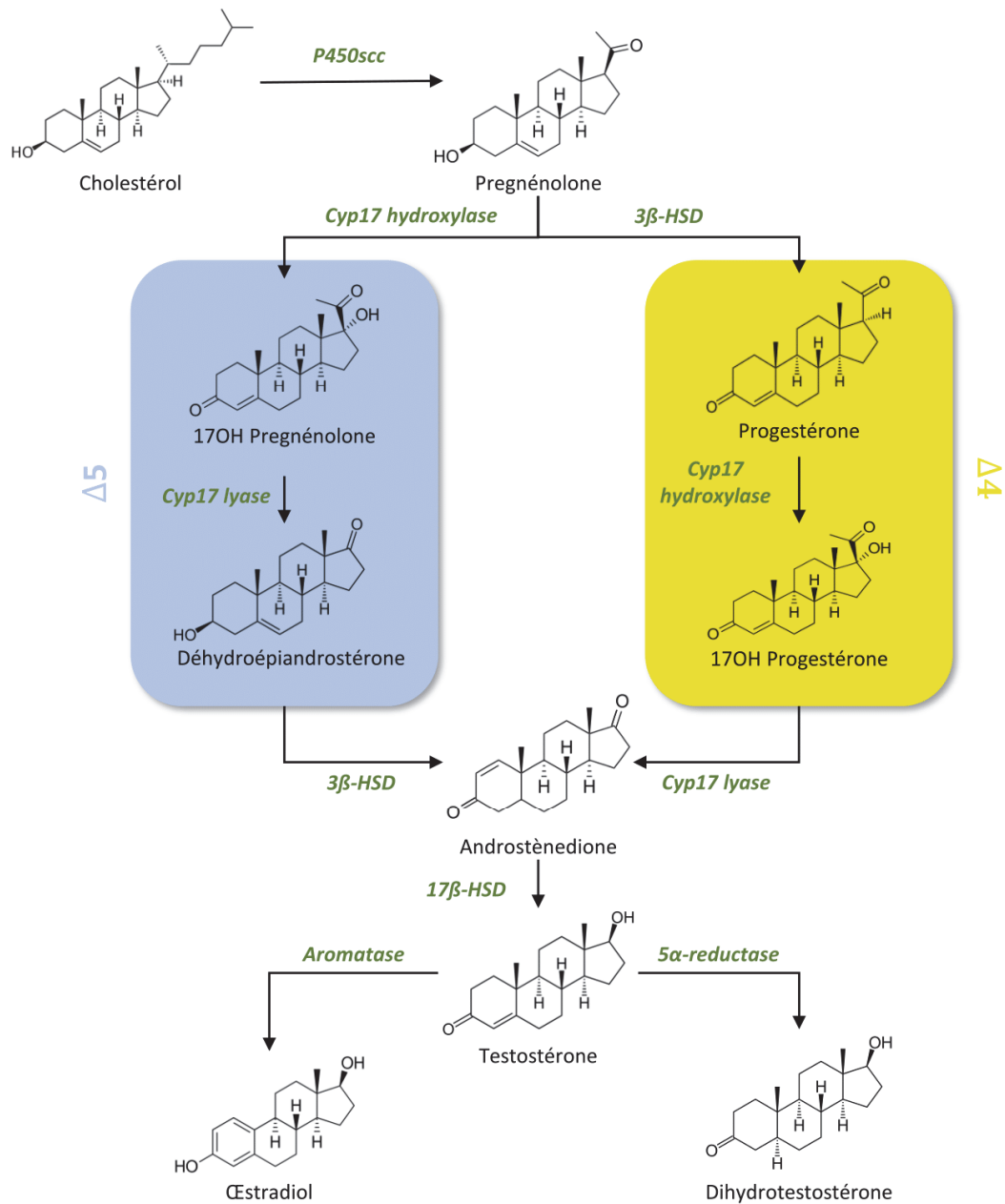


Figure 9. La stéroïdogénèse

Biosynthèse des stéroïdes sexuels chez l'homme dans les cellules de Leydig. La zone jaune indique les réactions catalysées par la 3 β HSD et sépare schématiquement la voie Δ^5 à gauche de la voie Δ^4 à droite. P450scc : enzyme de clivage de la chaîne latérale du cholestérol ; 3 β HSD : 3 β -hydroxystéroïde déshydrogénase/ Δ^5 - Δ^4 isomérase; 17 β HSD : 17 β -Hydroxystéroïde déshydrogénase.

1.2.3.3. Rôle des stéroïdes testiculaires

La testostérone se fixe sur les récepteurs aux androgènes (AR). Elle peut notamment agir localement au sein du testicule, sur les cellules de Sertoli, les cellules péritubulaires et les cellules de Leydig elles-mêmes, et réguler ainsi la spermatogénèse (Mainwaring, Haining, and Harper 1988). Elle joue également un rôle dans la différenciation, la croissance et le fonctionnement des organes reproducteurs. Cette hormone est à l'origine de l'acquisition des caractères sexuels secondaires et peut également influencer sur la distribution des graisses, contrôler la fonction érectile et avoir des effets psychotropes.

La dihydrotestostérone (DHT) se fixe sur les mêmes récepteurs que la testostérone, mais est impliquée dans des fonctions différentes du fait d'une plus grande affinité pour l'AR que la testostérone. La DHT est impliquée dans le développement des organes génitaux lors de la phase embryonnaire ainsi que dans la croissance de la prostate et la mise en place et le maintien de la pilosité. Elle est aussi capable d'induire une réponse plus rapide que la testostérone (Z. X. Zhou et al. 1995).

Les œstrogènes (l'œstradiol notamment) agissent via les récepteurs aux œstrogènes ER α et ER β exprimés au niveau des cellules de Leydig (ER α), de Sertoli (ER β) ainsi que dans les cellules germinales (Pelletier, Labrie, and Labrie 2000). Les œstrogènes modulent les effets des gonadotrophines et de la testostérone sur les fonctions testiculaires. Il a été démontré l'importance de ces hormones sur la spermatogénèse et la fertilité (O'Donnell et al. 1996). Enfin, les œstrogènes assurent le bon développement du système nerveux et sa masculinisation (Simpson 2000).

1.2.4. Mécanismes de régulation de la stéroïdogénèse

La régulation de la stéroïdogénèse s'effectue à trois niveaux différents. Le premier niveau s'effectue au niveau de l'hypothalamus par la sécrétion de GnRH, le deuxième niveau se situe dans l'hypophyse en charge de la synthèse de LH et FSH, et enfin le dernier niveau a lieu à l'intérieur même du testicule par des actions de rétrocontrôle de l'activine et de l'inhibine (Figure 10).

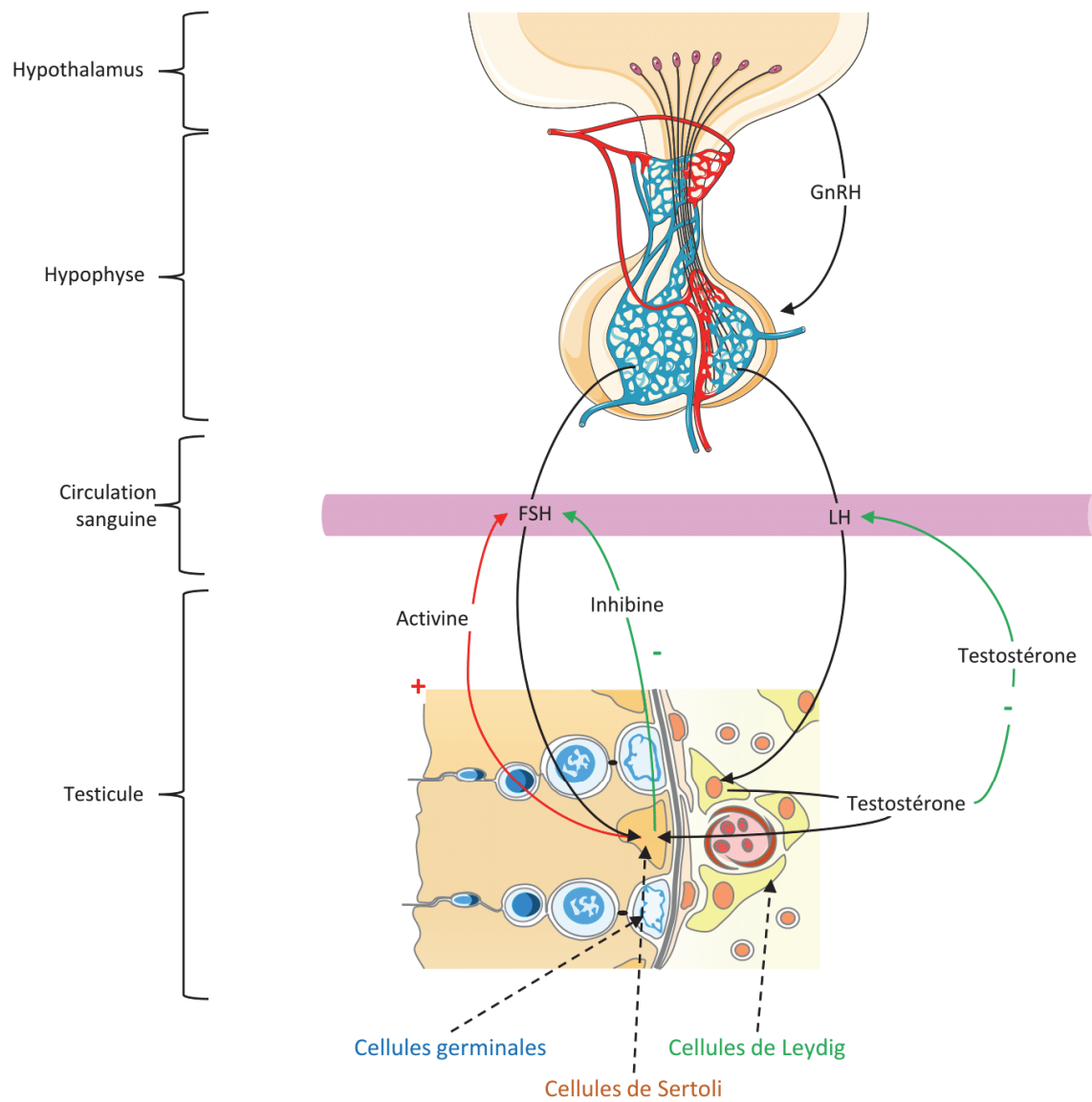


Figure 10. Mécanisme de régulation de la stéroïdogénèse

La régulation de la stéroïdogénèse s'effectue à trois niveaux différents. Le premier niveau s'effectue au niveau de l'hypothalamus par la sécrétion de GnRH, le deuxième niveau se situe dans l'hypophyse en charge de la synthèse de LH et FSH, et enfin le dernier niveau a lieu à l'intérieur même du testicule. La testostérone produite par les cellules de Leydig inhibe la production de LH et agit sur la cellule de Sertoli, produisant soit de l'activine (rétrocontrôle positif sur la LH) soit de l'inhibine (rétrocontrôle positif sur la LH).

1.2.4.1. L'axe hypothalamo-hypophyso-gonadique

La GnRH, neurohormone synthétisée au niveau de l'hypothalamus, est sécrétée de manière pulsatile. Elle est libérée par l'extrémité axonale des neurones hypothalamiques et cible les cellules gonadotropes de l'hypophyse. Sous l'action de la GnRH, deux hormones glycoprotéiques dimériques sont libérées dans la circulation générale : la LH et la FSH.

L'hormone lutéinisante ou LH, est l'activateur endogène principal du récepteur à la LH (LHR) présent au niveau des cellules de Leydig. Ce récepteur transmembranaire est couplé à une protéine G qui une fois activée stimule la production d'AMPc et par la même la stéroïdogénèse par la conversion du cholestérol en prégnénolone (Payne and Youngblood 1995). En parallèle de cette action rapide sur la stéroïdogénèse, la LH exerce une action locale de longue durée sur les cellules de Leydig entretenant le niveau d'activité enzymatique et les fonctions des autres organites impliqués dans la stéroïdogénèse.

L'hormone folliculo-stimulate ou FSH agit de manière indirecte sur la cellule de Leydig et donc sur la production de la testostérone. En effet, les cellules de Leydig ne possèdent pas de récepteurs à la FSH (FSHR). La FSH agit sur la cellule de Leydig par l'intermédiaire de facteurs protéiques paracrines sertoliens (IGF-1 ou le TGF- β).

1.2.4.2. Les rétrocontrôles

Les inhibines (A et B) sont des protéines dimériques capables de contrôler la FSH par une action directe sur l'hypophyse. L'inhibine B est principalement retrouvée chez l'homme et majoritairement produite par la cellule de Sertoli (de Kretser et al. 2001).

Les activines (A et B) sont des peptides dimériques. L'activine A, sécrétée par les cellules de Sertoli, les cellules péritubulaires et les cellules de Leydig durant la phase fœtale, exerce son effet sur la synthèse de la FSH produisant l'activine B au niveau hypophysaire (Makanji, Harrison, and Robertson 2011).

Les stéroïdes effectuent également un rétrocontrôle sur leur propre sécrétion en agissant sur la libération de la LH et la FSH au niveau de l'hypophyse et de l'hypothalamus (Franchimont, Chari, and Demoulin 1975). En effet, la testostérone, la dihydrotestostérone et l'œstradiol ralentissent les pulsations de GnRH hypothalamique, ralentissant ainsi la production de LH. La testostérone inhibe préférentiellement la LH, tandis que les œstrogènes inhibent la LH et la FSH.

1.3. Les perturbateurs endocriniens

Il est encore aujourd'hui relativement difficile de définir de façon exacte ce qu'est un PE (cf. 1.3.2. Un souci de définition). La définition la plus communément adoptée est celle de l'OMS énoncée en 2002, décrivant un PE comme « **une substance ou un mélange exogène, possédant des propriétés susceptibles d'induire une perturbation endocrinienne dans un organisme intact, chez ses descendants ou au sein de (sous)-populations.** ». À cette définition, l'OMS, ajoute un système de classement des PEs afin de les distinguer sous forme de trois catégories : Les PE avérés, suspectés et présumés. Bien que d'autres définitions soient employées par d'autres organismes et diffèrent de celle de l'OMS sur certains points, toutes, ou presque, s'accordent sur deux notions principales : la dérégulation du SE et l'induction d'un effet nocif pour la santé.

1.3.1. Histoire de la perturbation endocrinienne

Les premiers effets de perturbation endocrinienne rapportés dans la littérature scientifique datent des années 40. À cette époque, la notion de perturbation endocrinienne n'existe pas. Ces travaux mettaient simplement en évidence l'existence d'un effet délétère de l'environnement sur la reproduction des pygargues ou des visons aux États Unis (Wurster 2015).

Il faudra attendre 1962 pour qu'une première « sonnette d'alarme » soit tirée. Dans son livre « *Silent Spring* » traitant des conséquences du rejet de composés chimiques sur l'environnement, Rachel Carson met en évidence un impact direct de certaines substances sur la santé et la reproduction des espèces animales, dont l'Homme (Carson 1962). Elle s'est plus particulièrement intéressée aux effets d'un pesticide utilisé en grande quantité à cette époque, le dichlorodiphényltrichloroéthane ou DDT, sur la faune aviaire. Cette substance représente la première crise sanitaire liée à la notion de perturbation endocrinienne à grande échelle. Dans les années 60, 400'000 tonnes de DDT sont répandues dans le monde chaque année afin de protéger les cultures et d'augmenter leur rendement. Rachel Carson décrit comment le DDT est à l'origine de l'amincissement des coquilles d'œufs, contribuant ainsi à la diminution des populations aviaires étudiées. Par la suite, de nombreuses publications ont démontré la toxicité et notamment la reprotoxicité du DDT ainsi que d'autres pesticides tels que le méthoxychlore qui induit une féminisation des rats mâles (Tullner 1961). Sous l'impulsion de ces premières recherches d'évaluation du risque chimique environnemental, le premier symposium dédié aux œstrogènes dans l'environnement est organisé en 1979. Celui-ci a porté à la fois sur l'évaluation des propriétés chimiques et sur la diversité structurale de ces composés (McLachlan 1980).

Si les premières inquiétudes ont porté sur l'impact de produits chimiques sur la reproduction, le cas du diéthylstilbestrol (DES), puissant œstrogène de synthèse, a démontré que les produits de santé pouvaient également perturber le SE. Dans les années 1970, des millions de femmes enceintes dans le monde ont été traitées au DES, médicament délivré pour éviter les avortements spontanés à répétition. Des études épidémiologiques ont démontré que cette substance induisait des effets transgénérationnels (Klip et al. 2002). En effet, si l'effet du DES n'est pas visible sur les femmes traitées, leurs enfants (voire petits enfants) sont des populations à risque pour développer de nombreuses pathologies telles que des malformations de l'appareil génital, des cancers du sein ou bien une stérilité (Eskenazi et al. 2009). Ainsi, le DES s'est avéré être l'archétype du PE présentant des effets transgénérationnels aboutissant à son retrait du marché français en 1977.

Dans les années 1980 et 1990, de nombreuses publications sur l'impact des composés chimiques sur les fonctions de reproduction ont vu le jour, et ce partout dans le monde. En Angleterre, l'étude de *Sumpter et al.* met en évidence un changement de sexe chez des poissons exposés aux eaux rejetées par les stations de traitement des eaux usées. Les poissons males exposés développent un tissu ovarien et expriment des marqueurs œstrogène-dépendants (Sumpter and Jobling 1995). Aux États-Unis, c'est suite à la libération accidentelle de diclofol (composé proche du DDT) dans l'environnement du lac Apopka que de nombreuses études approfondies ont été menées par le professeur Louis Guillette. Ce spécialiste des alligators et de la perturbation endocrinienne a montré que l'exposition de ces animaux au diclofol entraînait une diminution du taux de testostérone avec pour conséquences une désorganisation testiculaire, le développement de micropénis et de fragments ovariens (Guillette et al. 1994). Plus tard, Theodora Colborn publie le livre « Our Stolen Future » (préfacé par le vice-président des États-Unis de l'époque, Al Gore) dans lequel il accuse les polluants chimiques d'être responsables, entre autres, de la baisse de la production spermatique, d'anomalies au niveau du tractus génital et de cancers. Ce livre pose également la question de savoir si, par la surutilisation de produits chimiques, l'humanité ne met pas en danger « sa fertilité, son intelligence et sa survie » (Colborn, Dumanoski, and Myers 1996). La Figure 11 présente la chronologie de l'intérêt porté aux PE.



Figure 11. Historique de la perturbation endocrinienne

Parallèlement à un nombre croissant de faits et de rapports alarmants (en rouge), la communauté scientifique et les organismes réglementaires ont mis en œuvre de nombreuses actions (en vert).

1.3.2. Un souci de définition

Comme expliqué précédemment, il est difficile de définir avec exactitude ce qu'est un PE. De la définition officielle découle plusieurs enjeux, notamment les aspects réglementaires concernant la mise sur le marché et la distribution de nombreux produits.

Le problème rencontré aujourd'hui avec les perturbateurs endocriniens est qu'il n'existe pas de définition officielle acceptée par tous les organismes, celle-ci évoluant au cours des années en fonction des études scientifiques menées. De plus, il n'existe aucun moyen de repérer toutes les classes de perturbateurs endocriniens compte tenu des nombreux modes d'action de ces composés. Une des premières définitions a été énoncée en 1996 par l'agence américaine de protection de l'environnement (US-EPA), selon laquelle un PE est « **un agent exogène qui interfère avec la synthèse, la sécrétion, le transport, la liaison, l'action ou l'élimination d'hormones naturelles responsables du maintien de l'homéostasie, de la reproduction, du développement et du comportement.** » (R J Kavlock et al. 1996). La même année, l'Union Européenne adopte sa propre définition à Weybridge en Angleterre. L'UE définit alors un PE comme « **une substance étrangère à l'organisme qui produit des effets délétères sur l'organisme ou sa descendance, à la suite d'une modification de la fonction hormonale** » (Commission 1996). Il est important de noter que cette dernière définition introduit le caractère délétère d'un PE. Ce concept instaure un débat problématique sur le fait qu'un PE est une substance qui a forcément un effet « négatif » sur l'organisme, et implique de définir d'une part, ce qu'est un effet négatif et d'autre part, sur quel organisme il s'exerce (uniquement l'Homme ? Le rat, la souris ... ?). Ainsi, très récemment, le 15 juin 2016, l'UE a présenté sa nouvelle définition des perturbateurs endocriniens : « **un perturbateur endocrinien est une substance ou un mélange exogène qui a un mode d'action endocrinien et qui est connu pour exercer des effets adverses pour la santé humaine sur un organisme intact, sa descendance ou ses sous populations comme conséquence de son effet mode d'action hormonale.** ». À peine énoncée, cette définition est déjà très décriée par de nombreux scientifiques. En effet, selon elle, un PE est une substance qui perturbe uniquement les populations humaines, et pour lequel il est nécessaire de mettre en évidence un lien de causalité entre les effets hormonaux et les effets délétères observés. Cette vision « binaire » de l'effet d'un PE implique la suppression des sous-catégories "suspectés" et "présumés".

L'OMS, quant à elle, décrit un PE comme « **une substance ou un mélange exogène, possédant des propriétés susceptibles d'induire une perturbation endocrinienne dans un organisme intact, chez ses descendants ou au sein de (sous)-populations.** ». C'est aujourd'hui la définition la plus communément utilisée. À partir de cette définition, l'Agence Nationale de Sécurité Sanitaire de

l'alimentation, de l'environnement et du travail (ANSES), en charge notamment du problème posé par les perturbateurs endocriniens, définit un PE comme « **tout produit chimique susceptible d'interagir directement avec le système endocrinien, et par voie de conséquence de produire un effet sur ce dernier et d'impacter les organes et les tissus.** »

Enfin, la dernière définition de perturbateurs endocriniens qu'il est nécessaire de présenter est celle exposée par la société d'endocrinologie américaine en 2012. Celle-ci apporte un nouveau regard sur ce qu'est un PE en éliminant le fait qu'un PE puisse avoir un effet délétère sur l'organisme. En effet, l'*Endocrine Society* présente les perturbateurs endocriniens comme des « **substances exogènes ou mélanges de substances pouvant interférer avec les mécanismes de l'action hormonale** » (R. Thomas Zoeller et al. 2012).

1.3.3. Mécanisme(s) d'action d'un perturbateur endocrinien

Les hormones sont la plupart du temps, produites à distance et relâchées au sein de la circulation sanguine. Elles sont transportées librement ou via un transporteur jusqu'à leur(s) cellule(s) cible(s) et se fixent sur un récepteur extra- ou intracellulaire de la cellule cible. La complexité des voies de signalisation hormonale rend ainsi possible une perturbation par un grand nombre de PE à différents niveaux et à différents endroits de l'organisme. Ceci rend difficile l'identification des PE et complexifie également la prédiction de leur(s) effet(s). À l'heure actuelle une grande partie des connaissances sur le mécanisme d'action des PE concerne leur capacité à interférer au niveau des récepteurs nucléaires hormonaux et de leurs transporteurs.

1.3.3.1. Effets hormono-mimétiques

Une substance hormono-mimétique est capable de mimer l'action d'une hormone endogène grâce à une forte similarité de structure moléculaire avec l'hormone (Figure 12A). La substance peut ainsi se lier au récepteur cible et induire les mêmes effets (effets dits agonistes). Les perturbateurs endocriniens les plus connus ayant un effet hormono-mimétique sont ceux à activité oestrogénique comme le DES ou le Chlordécone (Waring and Harris 2005).

1.3.3.2. Effets antagonistes

Comme pour les substances à effet hormono-mimétique, les substances ayant un effet antagoniste présentent une structure moléculaire similaire à l'hormone et se lient à son récepteur cible, mais sans l'activer (Figure 12B). La substance peut saturer les récepteurs de la cellule cible, bloquant ainsi le récepteur et les fonctions cellulaires associées comme la fixation du complexe hormone/récepteur

sur l'ADN. À titre d'exemple, le vinclozoline ou le flutamide présentent des effets antagonistes sur le récepteur des androgènes, et sont ainsi qualifiés de perturbateurs endocriniens anti-androgéniques (van Ravenzwaay et al. 2013; Gharagozlou et al. 2016).

1.3.3.3. Effets directs sur les hormones

Certaines substances n'altèrent pas le récepteur de l'hormone, mais agissent sur l'hormone elle-même (Figure 12C). Elles peuvent déréguler la synthèse de l'hormone, comme le kétoconazole qui inhibe les cytochromes P450 et par conséquent la synthèse de testostérone (Miossec, Archambeaud-Mouvier, and Teissier 1997). D'autres substances comme les phtalates bloquent le transport de l'hormone, en particulier les hormones lipophiles dépendantes de transporteurs pour atteindre leurs cellules cibles, comme la Sex Hormone-Binding Globulin (SHBG) (Sheikh et al. 2016).

1.3.4. Effets des perturbateurs endocriniens sur l'Homme

De nos jours, des composés potentiellement perturbateurs endocriniens sont détectables un peu partout dans notre environnement. De nombreuses études ont démontré que des expositions régulières, voire continues, à certains de ces composés sont responsables d'un nombre important de pathologies développées chez la personne exposée, mais également potentiellement chez sa descendance (Nassouri, Archambeaud, and Desailoud 2012). La littérature a en effet mis en évidence un grand nombre d'associations entre des expositions environnementales et des effets néfastes parmi lesquels des troubles du système nerveux, de la reproduction ou encore du développement embryonnaire. En 2013, l'Organisation mondiale de la Santé (WHO – *World Health Organization*) fait l'état des connaissances sur les PE ainsi que leurs effets chez l'Homme (A. Bergman, J. J. Heindel, S. Jobling, K. A. Kidd 2013). Face à la diversité d'action et des perturbations induites par ces PE, il ne sera évidemment pas possible d'être exhaustif dans la présentation de leurs effets.

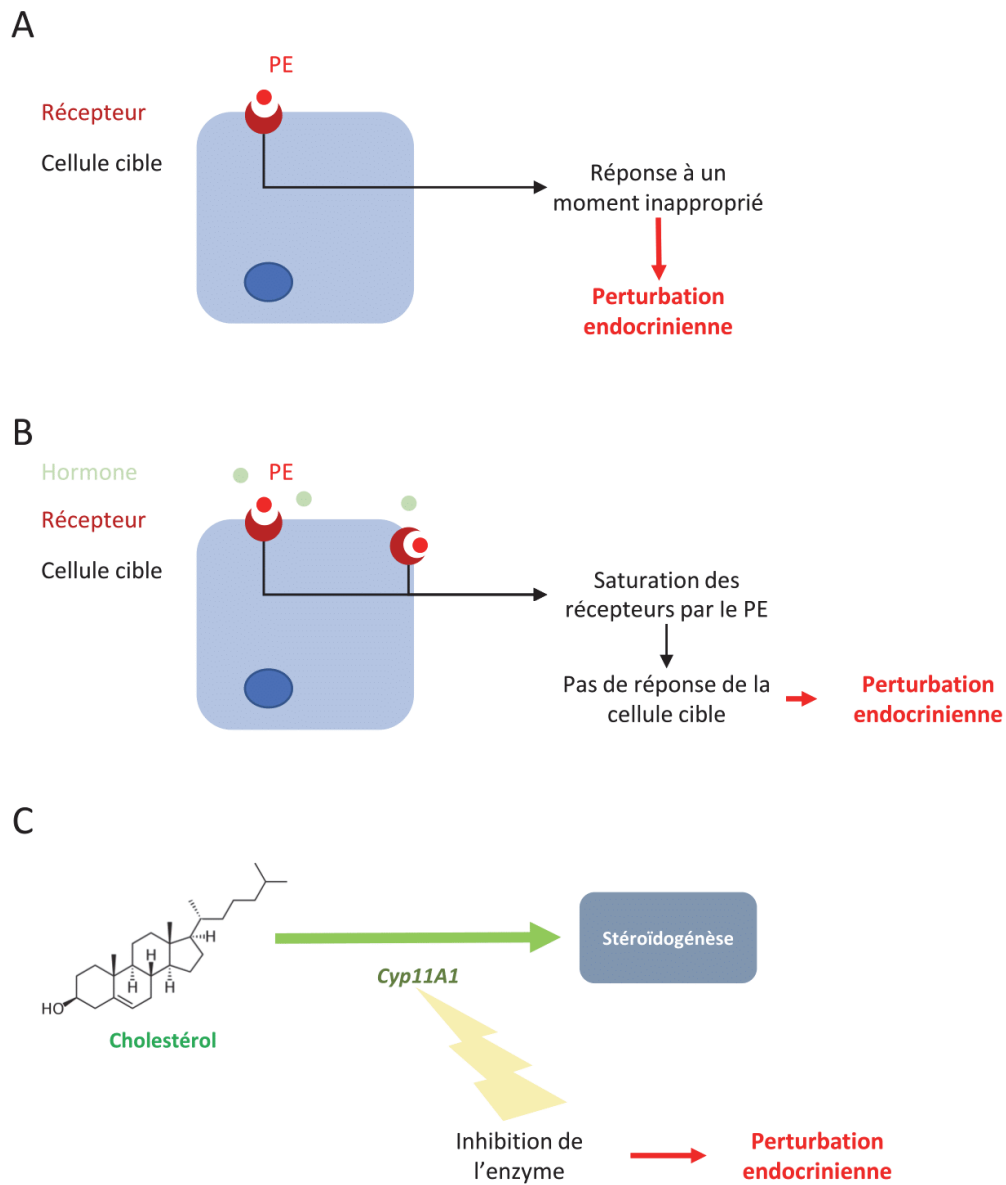


Figure 12. Différents modes d'action des PEs

A) Action hormono-mimétique du perturbateur endocrinien (PE). B) Action antagoniste du PE. C) Effet direct sur l'hormone via la perturbation d'enzyme de conversion (ici sur l'enzyme de clivage de la chaîne latérale du cholestérol, inhibant la testostérone).

1.3.4.1. Effets sur le système nerveux

En 2012, les grandes agences gouvernementales telles que l’OMS, l’Organisation des Nations Unies et le Programme National de Toxicologie aux États-Unis ont émis de fortes inquiétudes concernant l’effet des perturbateurs endocriniens sur le cerveau et le comportement (“WHO | Global Assessment of the State-of-the-Science of Endocrine Disruptors” 2013). Une étude sur les capacités cognitives des enfants exposés (ou dont les parents ont été exposés) à des perturbateurs endocriniens a mis en évidence une augmentation des troubles psychiatriques et neurologiques chez ces derniers (Boyle et al. 2011). Ils comprennent notamment des troubles du déficit de l’attention, des troubles autistiques, des syndromes dépressifs, des troubles de l’humeur, des difficultés d’apprentissage ou encore des troubles du comportement. Parmi ces perturbateurs endocriniens, les polychlorobiphényles (PCBs), utilisés comme isolants électriques, sont connus comme étant responsables du développement de troubles neurologiques chez l’Homme. Les métabolites des PCBs perturbent en effet le système thyroïdien et entraînent ainsi une augmentation du risque d’anomalies du développement du système nerveux (Winneke 2011), ayant pour conséquence une diminution du QI, des problèmes d’attention, de mémoire et de motricité fine comme l’écriture. Certaines des études apportant ces preuves ont été effectuées dans des communautés vivant à proximité de l’Arctique, un endroit longtemps considéré comme vierge, mais maintenant connu comme ayant bioconcentré les PCBs et d’autres polluants persistants au point d’atteindre, pour certains, les taux les plus élevés de la planète (Dewailly et al. 1993).

De même, les esters diphényliques polybromés (PBDEs) utilisés dans les retardateurs de flammes sont associés aux déficits cognitifs. Ils perturbent l’activité des neurotransmetteurs et des synapses, et par extension la viabilité neuronale (Dingemans, van den Berg, and Westerink 2011). Enfin des liens entre exposition aux pesticides et maladies neurodégénératives (comme la maladie de Parkinson) ou syndromes dépressifs ont aussi été démontrés (Freire and Koifman 2013).

1.3.4.2. Impact sur les cancers hormonodépendants

Le cancer du sein est le cancer le plus fréquemment retrouvé chez les femmes et est hormonodépendant dans 60 à 70% des cas. Des études ont mis en évidence des preuves indirectes de la forte relation statistique entre l’exposition aux PE et le risque de développement du cancer du sein. Cette hypothèse se trouve particulièrement renforcée par le suivi de cohortes de femmes ayant été exposées *in utero* au DES et qui présentent un risque plus élevé (multiplié par deux à 40 ans et par trois à 50 ans) de développer un cancer du sein (Palmer et al. 2006).

Chez l'homme de plus de 50 ans, c'est le cancer de la prostate qui est le cancer le plus fréquemment développé. Le risque de présenter un cancer de la prostate est à la fois positivement corrélé avec le taux d'estradiol, mais également négativement corrélé avec le taux de testostérone du patient. Aux Antilles il a été démontré une corrélation positive entre le taux sanguin de Chlordécone (pesticide organochloré oestrogénique ayant été massivement utilisé dans les bananeraies) et la survenue de ce cancer (Multigner et al. 2010).

1.3.4.3. Effets sur la thyroïde

Le développement du système nerveux est extrêmement dépendant des hormones thyroïdiennes pendant la période *in utero*, mais également durant les deux premières années de vie (R. T. Zoeller and Rovet 2004). Une altération de la production d'hormones thyroïdiennes maternelles ou fœtales, comme c'est le cas lors d'exposition au propylthiouracil, peut avoir de nombreuses complications telles qu'une hypothyroïdie fœtale, entraînant une réduction du nombre de synapses et de dendrites chez le nouveau-né (Gilbert et al. 2012).

1.3.4.4. Impact sur la fonction immunitaire

Récemment, une équipe de scientifiques a fait le lien entre l'effet potentiel des PE et les maladies auto-immunes. En effet, il a été montré que les femmes sont plus susceptibles que les hommes de développer des maladies auto-immunes telles que la sclérose en plaque, le lupus, la thyroïdite ou encore la myasthénie. Au sein du thymus (organe où sont formés les lymphocytes T), les lymphocytes sont exposés à des antigènes spécifiques des différent tissus (TSA) afin d'apprendre à reconnaître les cellules du soi. Cependant dans le cas des maladies auto-immunes, ces lymphocytes identifient les constituants normaux de l'organisme comme des cibles. Ceci est dû, notamment, à la diminution d'une protéine essentielle lors de la phase d'apprentissage des lymphocytes, la protéine AIRE pour AutoImmune REgulator, modulant le taux de TSA dans le thymus. Les œstrogènes ont un effet direct sur cette protéine et, en diminuant son expression, ils diminuent la capacité des lymphocytes à reconnaître les cellules du soi et augmentent ainsi la susceptibilité de l'individu à développer une maladie auto-immune (Dragin et al. 2016). Si rien n'a encore été démontré directement, il est licite de penser que les PE à activité oestrogénique ou anti-œstrogénique pourraient influencer sur la survenue de telles pathologies.

1.3.4.5. Modification du sex-ratio

Le sexe ratio à la naissance est le rapport entre le nombre de naissances de garçons et celui de filles. Dans le monde entier, ce ratio est assez constant avec environ 51,4% d'hommes pour 48,6 % de femmes (James 2004). Cependant, au cours de ces dernières années, une tendance à la diminution du nombre d'hommes est observée dans plusieurs pays industrialisés (Niels E Skakkebaek et al. 2016). Si les raisons de cette inversion ne sont pas claires, certaines études mettent en avant l'effet négatif des PEs sur la reproduction masculine (James 1995). L'exemple type de cette modification est celui de la catastrophe de Seveso. Le 10 juillet 1976, un nuage de dioxine (2, 3, 7, 8-tetrachlorodibenzo-para-dioxin ou TCDD) s'est échappé d'un réacteur d'une usine chimique se répandant dans les quatre communes avoisinantes. Une étude transgénérationnelle a démontré que, en plus de nombreux autres effets délétères, l'exposition au TCDD a complètement modifié le sex-ratio (12 filles pour 0 garçon) dans les familles exposées (Pesatori et al. 2003).

1.3.4.6. Altérations du métabolisme

L'augmentation du taux d'obésité dans le monde ne serait pas uniquement due au mode de vie et à l'alimentation. En effet, de plus en plus d'études suggèrent que d'autres facteurs comme l'exposition aux produits chimiques pourraient également jouer un rôle clé. Ces produits augmentant la prise de poids, dits obésogènes, modifient ou reprogramment des fonctions du SE contrôlant le métabolisme ou l'équilibre énergétique, aboutissant à des effets néfastes pour l'organisme en lien avec le surpoids (Grün and Blumberg 2007). Il a été démontré que l'exposition à ces molécules durant la vie fœtale et durant le plus jeune âge augmenterait les prédispositions au surpoids, mais également au diabète de type 2, aux maladies cardiovasculaires ou modifierait la sensibilité à l'insuline (Ismail-Beigi, Catalano, and Hanson 2006; Grün and Blumberg 2009).

1.3.4.7. Troubles de la reproduction chez la femme

L'interférence d'un PE avec les hormones ou leurs récepteurs peut être à l'origine de nombreux troubles reproductifs chez l'homme comme chez la femme. Plusieurs études ont notamment montré que les PEs peuvent être à l'origine d'une insuffisance ovarienne. L'insuffisance ovarienne primaire (POF) concerne environ 1% de la population féminine de moins de 40 ans et entraîne des troubles de la reproduction, des symptômes précoces de la ménopause et d'autres comorbidités (Nelson 2009). Si plusieurs causes ont été étudiées et peuvent être en lien avec les PEs, l'origine de la plupart des POF est inconnue, et probablement multifactorielle (Cooper et al. 2011).

D'autres études ont mis en évidence la capacité des PE à interférer dans le cycle menstruel, réduisant à terme la fécondité chez la femme. L'exposition fœtale et néonatale aux PE entraînerait, par interférence hormonale, une modification des cycles menstruels (Chao et al. 2007), par exemple une réduction de la durée des cycles menstruels après exposition aux pesticides organochlorés (Axmon et al. 2004). À l'inverse, les femmes exposées à des pesticides non organochlorés ont un risque de 60 à 100% plus élevé de présenter des cycles plus longs ou une absence de cycle (Farr et al. 2004).

Chez la femme il existe également de nombreuses interrogations sur la capacité des PE à induire une puberté précoce. Si cette hypothèse n'a pas encore été affirmée, de nombreuses études tendent à mettre en corrélation l'exposition aux composés chimiques tels que les phtalates ou les substances oestrogéniques avec ce phénomène (Buluş et al. 2016; Yum, Lee, and Kim 2013).

Enfin un trouble majeur fréquemment retrouvé chez la femme est le syndrome des ovaires polykystiques (PCOS) qui se caractérise par une anovulation chronique et un hyperandrogénisme. Alors que la pathogenèse du PCOS n'est toujours pas clairement établie, les données suggèrent que les facteurs génétiques, mais aussi environnementaux contribuent au développement de la maladie. Parmi les facteurs environnementaux connus, le bisphénol A (BPA) est un PE à action oestrogénique (Takeuchi et al. 2004).

1.3.4.8. Troubles de la reproduction chez l'homme

L'intérêt porté à la perturbation endocrinienne n'aurait sûrement pas été le même si leur découverte n'avait pas été accompagnée de la mise en évidence d'altérations de la fonction de reproduction, et notamment de la production spermatique.

En 1992, menée sous l'impulsion de Niels Erik Skakkebaek et Élisabeth Carlsen, une méta analyse de 61 articles publiés entre 1938 et 1990 regroupe les caractéristiques spermatiques de plus de 14900 hommes fertiles ou en bonne santé issus de tous les continents (E. Carlsen et al. 1992). Cette étude a mis en évidence la décroissance régulière de la production et de la densité spermatique au cours du temps, ainsi que du volume moyen de l'éjaculat. Bien qu'ayant agi comme un électrochoc au sein de la communauté scientifique, cette étude est vivement critiquée malgré la robustesse des méthodes statistiques employées. L'ensemble de ces altérations de la fonction reproductive a alerté la communauté scientifique et face à cela le chercheur danois Niels Erik Skakkebaek et ses collaborateurs ont formulé l'hypothèse que ces phénomènes pourraient être réunis et expliqués par une même entité : le Syndrome de Dysgénésie Testiculaire (TDS) (N E Skakkebaek, Rajpert-De Meyts, and Main 2001) (Figure 13).

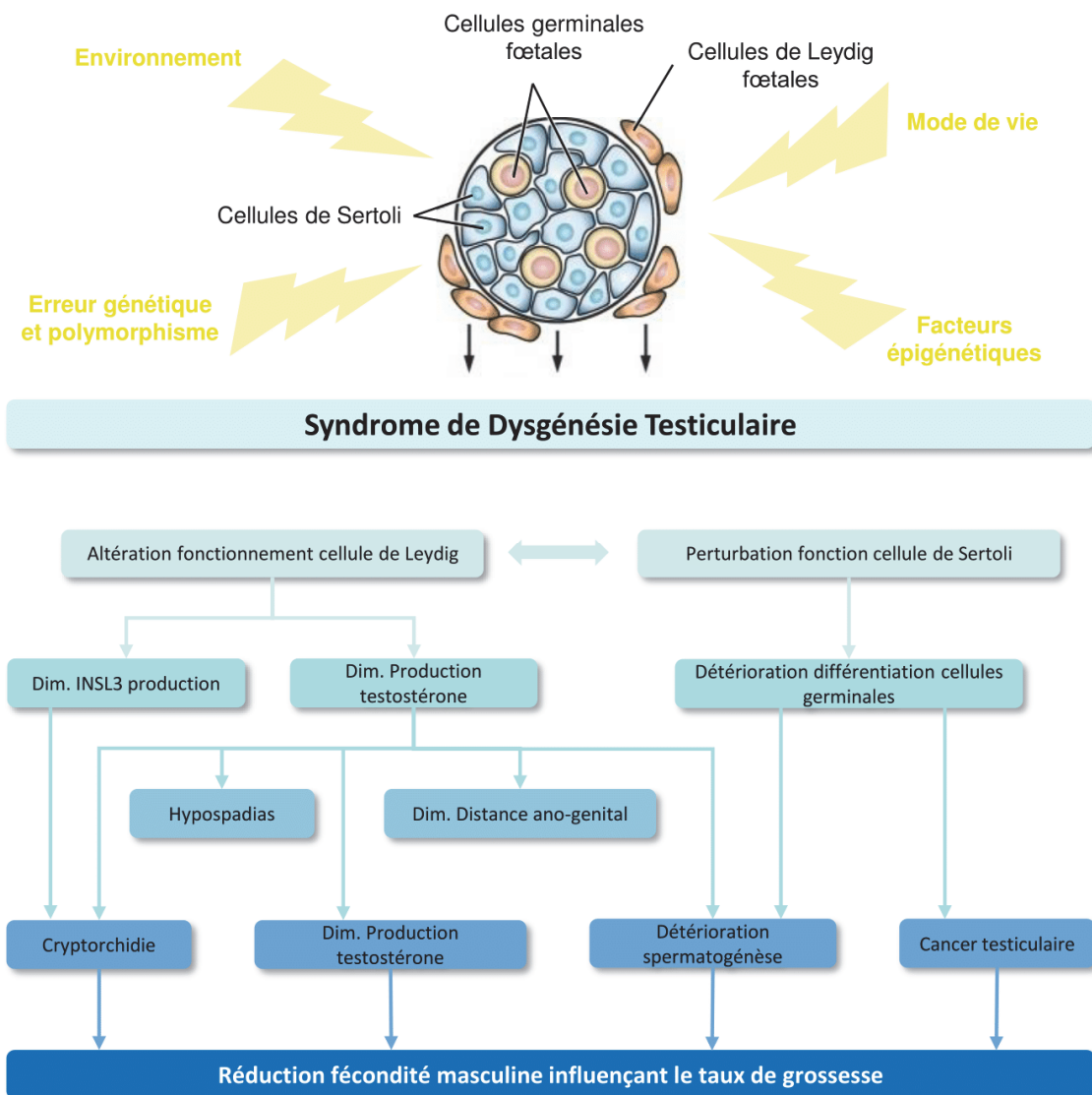


Figure 13. Schéma du syndrome de dysgénésie testiculaire selon Skakkebaek *et al.* 2016

Le chercheur danois Niels Erik Skakkebaek et ses collaborateurs ont formulé l'hypothèse que les différents phénomènes observés sur la reproduction chez l'homme pourraient être réunis et expliqués par une même entité : le Syndrome de Dysgénésie Testiculaire (TDS).

Il suggère ainsi que le déclin spermatique, l'augmentation des incidences de la cryptorchidie, de l'hypospadias et du cancer testiculaire, trouvent une origine commune pendant la vie fœtale par l'association d'altérations génétiques et d'expositions environnementales. Devant l'accumulation des indices liant ces quatre anomalies à la perturbation endocrinienne, l'hypothèse environnementale est la plus communément avancée (K M Main, Skakkebaek, and Toppari 2009), bien que n'étant pas unanimement admise. Si le TDS et son origine environnementale ne sont aujourd'hui qu'une hypothèse, il n'en demeure pas moins que de nombreuses études ont démontré que les PE peuvent être à l'origine de certains de ces phénomènes.

La cryptorchidie correspond à la non-descente de l'un ou des deux testicules dans le scrotum avant l'âge de 3 mois, et prend sa source lors du développement fœtal des gonades (Figure 14A). Les futurs testicules se développent dans l'abdomen et descendent au niveau du scrotum en deux phases: une phase transabdominale dépendante de l'INSL-3 puis une phase inguino-scrotale régulée par la CGRP (*Calcitonin Gene Related Peptide*) et la testostérone (Hutson et al. 2015). Ces deux phases doivent se produire durant le stade fœtal. Dans la plupart des cas de cryptorchidie, la descente survient dans les premiers mois de vie, compliquant ainsi le diagnostic. Si certaines études tendent à montrer une augmentation de la cryptorchidie dans la population, d'autres révèlent le contraire (Porter et al. 2005; Virtanen and Toppari 2007). On tente d'expliquer ce phénomène par l'omniprésence de PE, qui perturberaient le bon développement de l'appareil génital durant le stade fœtal : les herbicides (Mauduit et al. 2006), les fongicides (Norgil Damgaard et al. 2002), les phtalates (Parks et al. 2000) ou encore les insecticides (Weidner et al. 1998) sont notamment incriminés. D'autres études menées sur le testicule fœtal ont montré la présence d'une corrélation entre la prise d'antidouleurs chez la femme enceinte, comme l'ibuprofène (Ben Maamar et al. 2017), l'aspirine ou encore le paracétamol (Séverine Mazaud-Guittot et al. 2013), et le développement de cryptorchidie chez le nouveau-né. Si l'impact sur la diminution de la production d'INSL-3 est principalement mis en avant, l'effet de ces composés, et plus particulièrement du paracétamol, sur la production de testostérone est également mis en cause (van den Driesche et al. 2015). En 2015, Bernard Jégou dans une revue publiée dans le journal *Nature* faisait en particulier état des préoccupations actuelles vis-à-vis du paracétamol et de ses effets le testicule fœtal (pour revue, Bernard Jégou 2015).

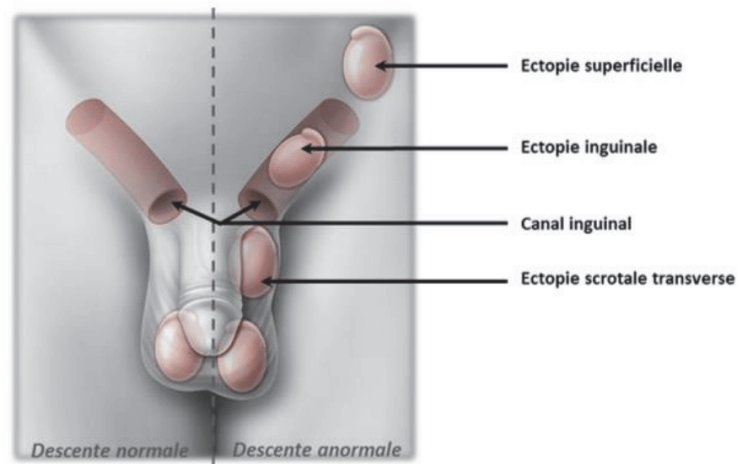
L'hypospadias est une anomalie congénitale de la fermeture de l'urètre (Figure 14B). Il se manifeste par une localisation anormale de son ouverture sur la surface ventrale de la verge ou au niveau du scrotum ou du périnée (L. Baskin 2017). Il constitue l'affection la plus fréquente du pénis, et son incidence se situe autour de 3%. Une étude a montré une augmentation de la prévalence de l'hypospadias

de 2.40% par an, et ce indépendamment de l'âge maternel (Lund et al. 2009). En 2010 et 2015, ces observations ont été remises en question par l'étude de données provenant des USA et du Royaume-Uni (Fisch, Hyun, and Hensle 2010) ainsi que de l'Europe (Bergman et al. 2015). Il a cependant été montré que l'exposition à un xénoestrogène ou à un xénoandrogène durant la période de développement fœtal peut interférer dans l'expression génique des récepteurs à androgènes ou des gènes *Hox*, essentiels dans l'organisation du système urogénital (L. S. Baskin, Himes, and Colborn 2001).

De nombreuses études ont essayé d'estimer l'impact des PE sur la qualité spermatique. S'il n'existe pas de consensus dans le domaine, certaines données laissent cependant penser que les PCBs, le BPA, les phtalates ou encore les dioxines (catastrophe de Seveso) peuvent être à l'origine d'une diminution de la concentration spermatique, de la mobilité des spermatozoïdes ou encore du nombre total de spermatozoïdes par éjaculat (Hauser et al. 2002, 2006; Knez et al. 2014; Mocarelli et al. 2008). Ces résultats sont nuancés par le fait que d'autres études n'ont pas pu mettre en évidence cette corrélation entre les PE et l'altération de la qualité spermatique (Jönsson et al. 2005; Meeker et al. 2010). Enfin, la cryptorchidie représentant un facteur de risque avéré d'une mauvaise qualité spermatique, il est possible que les PE, par induction de la cryptorchidie (cf. ci-dessus), puissent être indirectement responsables de mauvais paramètres spermatiques.

Le cancer du testicule est un cancer rare ne représentant qu'1 à 2% des cas de cancer chez l'homme, mais il constitue le premier cancer de l'homme jeune, entre 15 et 35 ans. Depuis plus de 50 ans, l'incidence du cancer du testicule est en forte et constante augmentation dans la plupart des pays industrialisés (Chia et al. 2010; Huyghe, Plante, and Thonneau 2007). Plusieurs facteurs de risque ont à ce jour été mis en évidence, notamment la cryptorchidie (Gurney et al. 2017). Son apparition est également plus fréquente en cas d'antécédent familial de cryptorchidie, de cancer du testicule, de la prostate, ou du sein (Walschaerts et al. 2007). Malgré ces facteurs, des nombreuses questions subsistent quant à l'existence d'autres origines de la maladie, et plus précisément sur le rôle éventuel d'expositions environnementales.

A



B

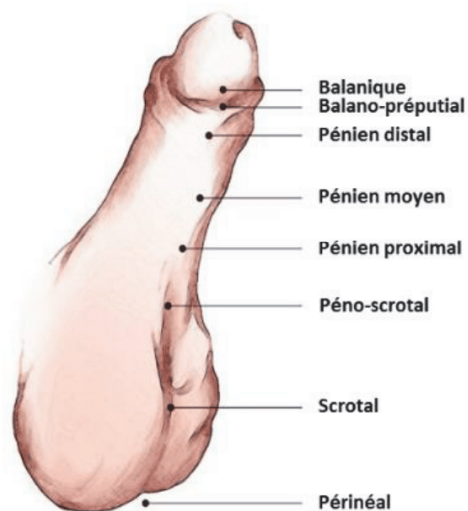


Figure 14. Schéma représentatif de la cryptorchidie et de l'hypospadias

A- La cryptorchidie correspond à la non-descente de l'un ou des deux testicules dans le scrotum. B- L'hypospadias est une anomalie congénitale de la fermeture de l'urètre.

1.3.5. Présentation de perturbateurs endocriniens connus

L'identification des substances pouvant agir comme des PE est une priorité pour les agences de régulation. Cependant, cette tâche n'est pas aisée : environ un millier de composés ont, sur la base d'informations scientifiques, été jusqu'alors démontrés comme perturbant les voies de signalisation hormonale. Certaines molécules ont fait l'objet d'investigations poussées et de réglementations strictes, voire même ont été retirées du marché. Cependant, pour la plupart de ces substances, il n'existe que des informations fragmentaires empêchant de statuer sur leurs effets sur le SE. Les molécules utilisées peuvent être transformées en de nombreux produits de dégradation, et de ce fait impacter indirectement le fonctionnement du système hormonal. De manière générale, il existe trois grands groupes de perturbateurs endocriniens : les produits naturellement présents dans l'environnement, les produits de synthèse fabriqués pour un usage ciblé et les substances de synthèse présentant l'activité hormonale recherchée. Parmi les composés de synthèse, trois grandes familles se détachent : les polluants organiques persistants (POPs), les pesticides et les plastifiants.

1.3.5.1. Les substances naturelles

Un certain nombre PE connu est d'origine naturelle et produits par les végétaux ou les champignons pour ne citer qu'eux. Les phytoœstrogènes sont des molécules issues du règne végétal capables de se fixer sur le récepteur aux œstrogènes. L'identification de troubles de la reproduction comme l'infertilité dans des troupeaux ovins australiens (dite "maladie du trèfle") dans les années 40 a mis en avant l'impact de ces phytoœstrogènes (Adams 1990). Parmi eux, on peut citer la génistéine, la daidzéine ou le coumestrol. Ils sont présents dans certaines plantes ou légumes comme le trèfle, les germes de soja, la luzerne, les choux, les haricots verts et les épinards (Boué et al. 2003). À ce jour plus de 200 substances issues de plantes ont été reconnues comme ayant des activités oestrogéniques. La majorité des phytoœstrogènes appartiennent au groupe des flavonoïdes. Ce dernier est divisé en trois catégories principales : (i) les isoflavones présentes dans les légumineuses ; (ii) les lignanes sont retrouvés dans diverses céréales ; (iii) les coumestanes, essentiellement représentés par le coumestrol, présents dans les luzernes utilisées en alimentation animale.

Les mycoœstrogènes sont des substances produites par les champignons. Ils sont retrouvés à de fortes concentrations après la récolte du maïs ou de l'orge, ainsi que dans les graines de céréales et dans l'huile végétale. On distingue parmi les mycotoxines, la zéaralénone. Cette dernière se développe dans les denrées alimentaires à la suite d'une infection fongique des céréales notamment le maïs, mais aussi

l'avoine, le blé, le riz et le soja (Massart and Saggese 2010). L'effet toxique le plus préoccupant de la zéaralénone est son caractère de PE à activité oestrogénique (AFSSA 2009).

1.3.5.2. Les polluants organiques persistants

Les polluants organiques persistants (POPs) sont des composés organiques qui présentent des résistances aux dégradations photolytiques, biologiques et chimiques. Il n'est donc pas rare de retrouver dans l'environnement, des années après leurs utilisations, ce type de composés. Les POPs sont en majorité des molécules halogénées, qui se caractérisent par une faible solubilité dans l'eau et une forte solubilité dans les lipides, entraînant ainsi une bioaccumulation dans les tissus adipeux (K. C. Jones and de Voogt 1999; Lee et al. 2017). Nombre de ces composés ont été, ou continuent d'être, utilisés en grandes quantités. En raison de leur persistance dans l'environnement, ils peuvent être bioaccumulés et bioamplifiés. La bioaccumulation est le mécanisme selon lequel des substances chimiques absorbées par des organismes vivants vont s'accumuler dans certains tissus (adipeux la plupart du temps). La bioamplification est le mécanisme par lequel la teneur en substances chimiques est augmentée tout au long de la chaîne alimentaire. Du fait de cette bioamplification, certains organismes sont capables de concentrer jusqu'à 1 million de fois les substances toxiques (moules, huîtres) (Luna-Acosta et al. 2015) et donc de devenir toxiques. Cette bioaccumulation pousse aujourd'hui les autorités sanitaires à mettre en garde les femmes enceintes vis-à-vis de la consommation de certains animaux, comme les crustacés ou les poissons sauvages. La prise de conscience sur les dangers suscités par ces POPs a poussé 178 pays et l'Union Européenne à ratifier un accord, la convention de Stockholm, entrée en vigueur en 2004 (Hung, Katsoyiannis, and Guardans 2016). L'objectif de cette convention est de réduire, voire d'éliminer, la production et l'utilisation des POPs.

Parmi les POPs les plus connus, les dioxines et les furanes sont des composés polychlorés présentant des structures moléculaires et un mode d'action similaires. En Europe, la principale source de dioxine provient des incinérateurs de déchets, qui lors de la combustion relâchent les dioxines (sous-produits de la combustion). Chez l'Homme, l'exposition s'effectue presque entièrement par l'alimentation, et plus particulièrement par les produits laitiers, le poisson et la viande. Les dioxines sont des composés lipophiles métabolisés et éliminés lentement par l'organisme, tendant donc à la bioaccumulation. La dioxine la plus connue est le 2,3,7,8-tétrachlorodibenzo-p-dioxine ou TCDD, classé comme cancérigène depuis 1997. Il a été prouvé que le TCDD est responsable de l'induction d'une enzyme de détoxification, de la modulation de l'expression de gènes, d'une hépatotoxicité et de l'altération des fonctions de reproduction et du développement (Mandal 2005). Une exposition périnatale au TCDD aurait des effets sur l'appareil reproducteur mâle (L. Gray et al. 1995). Ces effets incluent un

retard pubertaire, une diminution de la concentration plasmatique de testostérone, une diminution de la production spermatique et une diminution du poids des organes sexuels secondaires. L'ensemble de ces effets ressemble à ceux induits par des anti-androgènes tels que la vinclozoline ou les phtalates, et il a été montré qu'ils seraient dus à une activité anti-androgénique passant par l'Ahr (Recepteur Aryl hydrocarbure) (Safe 2001).

1.3.5.3. Les pesticides

Une classe autre que les POPs fait aujourd'hui l'objet de nombreuses régulations : les pesticides. Ces substances ont été mises sur le marché depuis plusieurs décennies à un rythme soutenu, générant de nombreux problèmes sanitaires. Ces pesticides ont été, et sont toujours, déversés en masse dans l'environnement, provoquant une contamination généralisée des sols et nappes phréatiques. Ils sont également responsables de nombreux troubles de la santé chez l'Homme. Si le caractère néfaste des pesticides est bien documenté, il n'en reste pas moins difficile de supprimer ces substances du marché. En effet, ces substances sont sujettes à un fort lobbying industriel cherchant à préserver ce marché de plus 40 milliards de dollars (Grube et al. 2006). Une partie du problème posé par les pesticides est que ces derniers ne restent pas exclusivement dans les sols où ils sont épandus. À cause de l'évaporation, une grande partie de ces substances est dispersée dans l'atmosphère. De plus, en 2005, une étude de l'INRA a mis en évidence que 50 à 90% des pesticides épandus n'atteignaient pas leur cible, à savoir les nuisibles animaux ou végétaux.

Parmi ces pesticides, le Chlordécone est un pesticide, mais également un POP de la famille des organochlorés, interdit dans de nombreux pays. Il possède une demi-vie (i.e. la durée à l'issue de laquelle sa concentration initiale dans le sol est réduite de moitié) comprise entre 120 et 160 jours dans l'organisme et pouvant aller jusqu'à 46 ans dans l'environnement (Multigner et al. 2016). Le Chlordécone a été massivement utilisé aux Antilles françaises, pour la lutte contre le charançon du bananier entre les années 70 et 90. De par sa biopersistance dans les nappes phréatiques et les sols, une exposition des populations est toujours visible. Dans le milieu des années 1970, les premiers signes de toxicité du Chlordécone ont été mis en évidence : les personnes exposées présentaient les symptômes cliniques d'un empoisonnement chronique ciblant le système nerveux (céphalées, troubles oculomoteurs, pertes de mémoire), une hypertrophie hépatique, et des troubles touchant le système reproducteur (diminution de la fraction de spermatozoïdes motiles) (Cannon et al. 1978; Faroon et al. 1995; Taylor 1985). Aujourd'hui les effets du Chlordécone sont bien documentés et il est possible de les lister (Multigner et al. 2016). Cette substance se fixe sur les récepteurs à ostéogène stimulant la synthèse de progestérone, hormone stéroïdienne jouant un rôle clé dans le maintien de la grossesse, entraînant ainsi

une naissance prématurée. De plus, la capacité du Chlordécone à traverser la barrière placentaire perturbe le neurodéveloppement lors de la vie fœtale. Enfin l'effet agoniste du Chlordécone sur les récepteurs à œstrogènes α , responsable d'inflammations et de prolifération cellulaire aberrante au sein de la prostate, est à l'origine du développement de cancer de la prostate.

Le glyphosate, molécule active du Roundup, est l'herbicide le plus utilisé au monde. Il est produit par la société Monsanto depuis 1974 et est aujourd'hui le pesticide le plus controversé aussi bien médiatiquement que scientifiquement. Le glyphosate est une amine méthylphosphorylée. Bien que dégradable dans l'environnement, la surutilisation de ce pesticide en fait un polluant systématiquement retrouvé dans les écosystèmes. Des études ont mis en évidence la présence de glyphosate dans des poussières ou aérosols issus d'activités agricoles. Ces poussières, véhiculées par le vent, font de l'inhalation l'une des sources potentielles d'exposition à ce pesticide (Grunewald et al. 2001). La principale controverse autour du glyphosate vient de ses effets sur la santé humaine. Ce débat a été rendu public pour la première fois suite à la publication de Séralini *et al.* en 2012 (Séralini et al. 2012) qui a mis en évidence la modification de l'équilibre hormonal ainsi que le développement de tumeurs mammaires chez le rat après exposition au Roundup. Sujet à de nombreuses critiques, cet article fut par la suite rétracté. Ce travail a néanmoins permis d'alerter la communauté scientifique, mais également les médias sur les potentiels dangers de cet herbicide. Par la suite, il a été démontré que le glyphosate pouvait avoir des effets sur les organes reproducteurs de par son effet anti androgénique (Mesnage et al. 2015). À l'heure actuelle, le glyphosate est toujours autorisé sur le marché. En juin 2016, alors que son innocuité pour la santé humaine est actuellement remise en cause, son autorisation d'utilisation a été prolongée de 18 mois par la commission européenne, afin d'attendre que l'Agence Européenne des Produits Chimique (ECHA) émette son avis au plus tard fin 2017. Enfin, en mars 2015, le Centre International de Recherche sur le Cancer (CIRC) classait le glyphosate comme cancérogène probable, alors qu'en novembre 2015 l'EFSA, l'European Food Safety Authority, a estimé qu'il était improbable que ce pesticide soit un cancérogène pour l'Homme (Portier et al. 2016).

Un autre herbicide très utilisé en France jusqu'à son interdiction en 2003 est l'atrazine. Cette substance est un herbicide de la famille des triazines interférant avec les mécanismes de photosynthèse. Cette molécule a une demi-vie estimée de quelques semaines à plusieurs mois (Solomon et al. 2008). L'atrazine se retrouve dans les eaux de surface et, dans une moindre mesure, dans les nappes d'eau souterraines, ainsi que dans les sols de pays où son utilisation a été interdite (Vonberg et al. 2014). Bien qu'aujourd'hui interdit en France, cet herbicide est massivement utilisé aux États Unis pour traiter les champs de maïs à raison de plus de 400.000 tonnes par an. De par son importante utilisation, des inquiétudes ont été émises vis-à-vis de ses effets sur l'environnement et la santé humaine (Gammon et

al. 2005). Il a été démontré que ce pesticide est capable d'interférer avec les fonctions de phosphorylation oxydative (processus de transformation de l'ADP en ATP) ainsi que les fonctions des cytochromes P450, diminuant la consommation d'oxygène de la cellule (Bhatti, Sidhu, and Bhatti 2011; Jin et al. 2014). Cette molécule altère également le processus de reproduction chez le rat (Victor-Costa et al. 2010), dérégule le développement neurologique chez le poisson-zèbre (Wirbisky et al. 2015), baisse le niveau d'androgènes et est responsable de problèmes de différenciation sexuelle chez les amphibiens (T. B. Hayes et al. 2002; T. Hayes et al. 2003; T. B. Hayes et al. 2010). Comme pour le glyphosate, ces études ont été très controversées, mais de nombreuses interrogations persistent. Enfin, une étude épidémiologique sur la cohorte PELAGIE a montré qu'il existait une corrélation entre le taux d'atrazine retrouvé dans les urines et le développement embryonnaire (diminution du poids de naissance ou périmètre crânien) (Chevrier et al. 2011). En parallèle, une autre étude a montré l'impact de l'atrazine sur l'épigénome, ainsi que sur la diminution de la production de testostérone et de spermatozoïdes (Gely-Pernot et al. 2015).

1.3.5.4. Les plastifiants

Chaque année, près de 9 millions de tonnes de plastifiants sont utilisées afin d'améliorer la solidité, l'aspect et la souplesse des plastiques. Les plus connus, le bisphénol A (BPA) et les phtalates sont au centre de nombreuses interrogations scientifiques et médiatiques. En effet, en quarante ans, les plastiques ont envahi notre environnement. Leur omniprésence pose un grand nombre de questions sur l'environnement et la santé publique, notamment en termes de perturbation endocrinienne.

Le BPA est le plus connu des bisphénols et des plastifiants en règle générale (pour revue, Rubin 2011). Il est très utilisé dans l'industrie comme monomère et précurseur dans la fabrication de résines époxydes (revêtement interne de boîtes de conserve) et de polycarbonates, révélateur d'encre d'impression thermique (tickets de caisse), ou encore de billets de banque. L'un des problèmes soulevés par cette molécule est qu'une fois chauffée ou exposée à des détergents, elle est capable de migrer des plastiques alimentaires vers les aliments, pouvant ainsi être ingérée. Bien que l'ingestion soit la voie d'exposition principale, l'exposition au BPA peut également être transcutanée (Zalko et al. 2011). Le BPA se retrouve un peu partout dans le monde, si bien que deux études, l'une américaine (Calafat et al. 2008) et l'autre canadienne (Vandenberg 2011), ont montré que plus de 90% des échantillons d'urine collectés contenaient du BPA en quantité détectable, traduisant l'exposition omniconstante de la population à cette molécule. Le BPA est un PE connu, interdit en France depuis 2010 dans les biberons et les contenants alimentaires. Ce plastifiant interfère avec les œstrogènes, entraînant une diminution de la spermatogénèse (Chevrier et al. 2011). Il est également associé à des effets néfastes sur le

développement de l'appareil reproducteur masculin entraînant une augmentation de l'infertilité et du nombre de cancers (Maffini et al. 2006), ainsi que des anomalies du développement prostatique (Timms et al. 2005). Le BPA possède également un effet anti-androgénique sur les testicules durant la vie fœtale (Maamar et al. 2015) et adulte (Desdoits-Lethimonier et al. 2017). Les effets délétères de ce plastifiant étant de plus en plus connus, les gouvernements ont pris des mesures visant à l'interdire ou à en limiter l'utilisation. Ces changements dans la réglementation ont poussé les industriels à chercher des alternatives, notamment par la création d'analogues dénommés substituts du BPA, pour lesquels l'innocuité n'est cependant pas garantie non plus, notamment concernant le testicule (Eladak et al. 2015).

Le deuxième grand groupe de composés plastifiants préoccupants est constitué des phtalates (pour revue, Zlatnik 2016), tels que le di(n-butyl)phthalate (DBP) ou le di(2-éthylhexyl)phthalate (DEHP). Depuis les années trente, ces composés sont utilisés comme plastifiants et entrent dans la composition de nombreux produits tels que les cosmétiques, jouets pour enfant, peintures ou encore emballages alimentaires. Comme pour le BPA, la voie principale d'exposition est l'ingestion suite à la migration de l'emballage vers l'aliment. De plus, il se pourrait que ces substances puissent être ingérées par les enfants lorsque ceux-ci portent leurs jouets à la bouche. De fortes concentrations en métabolites provenant des phtalates ont été détectées dans les urines de femmes en âge d'avoir des enfants provenant de l'usage d'objets du quotidien (Schettler 2006; Blount et al. 2000). Les phtalates possèderaient également une action anti-androgénique. Il est actuellement supposé qu'au lieu de se lier aux récepteurs aux androgènes, ces molécules induiraient une diminution de la synthèse de testostérone des cellules de Leydig, diminuant ainsi le taux de testostérone durant la vie fœtale, période critique de la différenciation sexuelle (Parks et al. 2000). Il a été montré que l'exposition au cours de la différenciation sexuelle entraîne de nombreux effets sur le développement de l'appareil reproducteur mâle chez le rat, comme la diminution de la distance anogénitale ou encore une rétention des aréoles et des mamelons, une augmentation de cryptorchidie et d'hypospadias, une diminution de la production spermatique, une hyperplasie des cellules de Leydig ou encore une diminution du poids prostatique (pour revue, Albert & Jégou 2014; Gray et al. 2000). De plus, il a été démontré une association entre l'exposition au DEHP et la diminution de la largeur pénienne (Swan 2008) et le ratio d'hormones chez le nouveau-né (Katharina M Main et al. 2006). De nombreux pays ont ainsi restreint ou interdit les importations de produits contenant du DEHP, encourageant les industriels à rechercher des alternatives moins risquées pour la santé humaine.

1.3.5.5. Les médicaments

Récemment, de plus en plus d'études mettent en garde contre la consommation de certaines substances à portée médicinale. Si parmi ces substances il est normal de citer les pilules contraceptives, il n'est pas forcément intuitif de penser également à l'aspirine ou encore au paracétamol.

La pilule contraceptive est essentiellement basée sur l'utilisation d'œstrogènes ou de progestagènes contrôlant les cycles ovariens chez la femme. Ces molécules bloquent l'ovulation et provoquent un épaissement de la glaire cervicale qui empêche toute nidification d'œuf fécondé. Au vu des molécules utilisées au sein de ces pilules (éthinylestriol, mestranol ...) et de leur mode d'action, il s'agit ici d'un exemple de perturbation endocrinienne. En effet ces œstrogènes vont venir bloquer l'ovulation par rétrocontrôle sur l'axe hypothalamo-hypophysaire empêchant la production de GnRH et ainsi de LH et de FSH.

Parmi les médicaments les plus utilisés figurent le paracétamol, dont la consommation mondiale ne cesse d'augmenter, et les AINS (anti-inflammatoires non stéroïdiens) comme l'aspirine. Une étude menée en 2013 sur ces deux molécules a mis en avant leur effet PE chez l'homme (O Albert et al. 2013). Après exposition de testicules adultes à ces substances dans des doses similaires à celles retrouvées dans le plasma sanguin, Albert et ses collaborateurs se sont aperçus d'une modification de la production de testostérone et d'INS β -3 par les cellules de Leydig. De plus, comme exposé précédemment (cf. 1.3.4. Effets des perturbateurs endocriniens sur l'Homme) ces médicaments sont à l'origine du développement de cryptorchidie chez le nouveau-né après exposition de la femme enceinte. C'est pourquoi il est fortement déconseillé à ces femmes de prendre de l'ibuprofène, mais également du paracétamol durant le dernier trimestre de la grossesse.

1.4. Évaluation des risques

L'objectif de la toxicologie est d'étudier les effets néfastes d'agents chimiques, biologiques ou physiques sur les systèmes biologiques. Il est évident, au vu de l'impact des PEs sur l'organisme qu'une détection rapide de leurs effets et de leurs mécanismes d'actions est nécessaire afin d'éviter de répéter les erreurs du passé (cf. 1.3.1. Histoire de la perturbation endocrinienne). Toutefois, un certain nombre de caractéristiques des PEs et de leurs modes d'action rendent peu adaptés les outils classiques de toxicologie. Aussi il a été nécessaire de repenser la manière d'identifier les effets néfastes de ces molécules (Harvey and Everett 2006). Cela passe, bien évidemment par le développement de nouveaux tests de détection, mais également par l'utilisation des nouvelles technologies.

1.4.1. La réglementation européenne

Depuis 2007, l'Union Européenne a adopté la réglementation REACH. Cette réglementation impose l'enregistrement et l'évaluation de la toxicité de toutes nouvelles substances chimiques avant leur mise sur le marché. L'objectif principal de cette réglementation est d'améliorer les connaissances et les caractéristiques des substances chimiques. Sont concernés par REACH tous les composés nouvellement mis sur le marché, mais également toutes les molécules mises sur le marché depuis 1981. Pour cela, il est aujourd'hui obligatoire d'enregistrer chaque composé chimique produit ou importé dans l'UE à raison de plus d'une tonne par an. Cet enregistrement implique la création d'une fiche de sécurité, « carte d'identité » de la substance regroupant toutes les informations relatives à sa toxicité et son écotoxicité. Cette fiche sera ensuite évaluée par l'Agence Européenne des produits CHimiques (ECHA) qui interdira ou classera la substance selon trois niveaux : 1) la molécule est classée sans risque et peut être mise sur le marché ; 2) la molécule présente des risques maîtrisables, son utilisation est donc réglementée ; 3) les risques engendrés par la molécule sont jugés trop importants, elle est interdite et doit être substituée par une autre molécule. La Figure 15 présente le processus d'évaluation des risques selon REACH d'après Penman et ses collaborateurs (Penman et al. 2015).

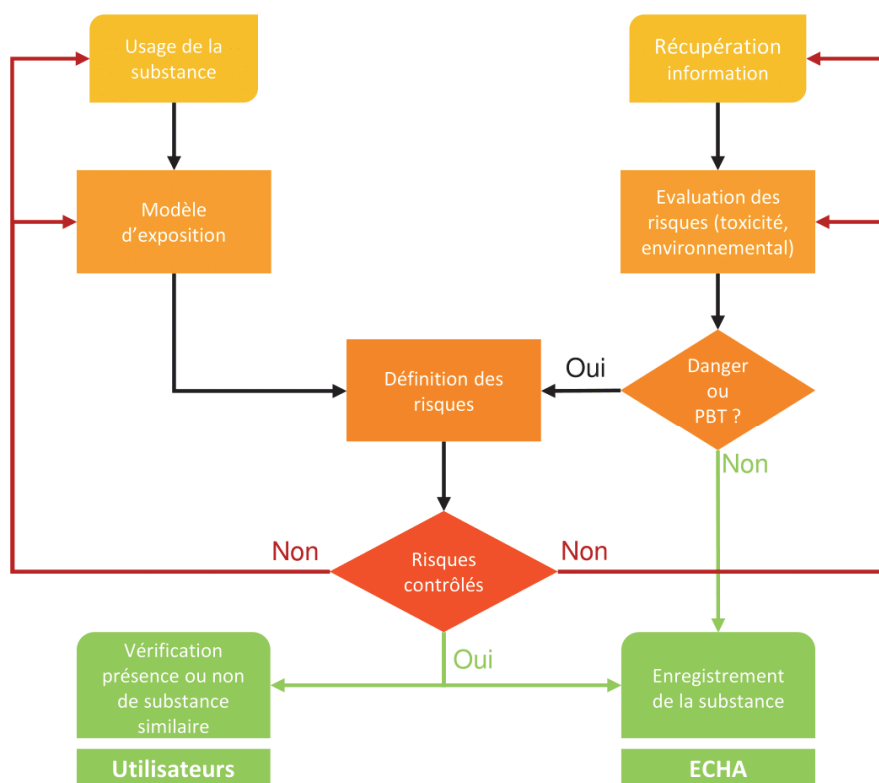
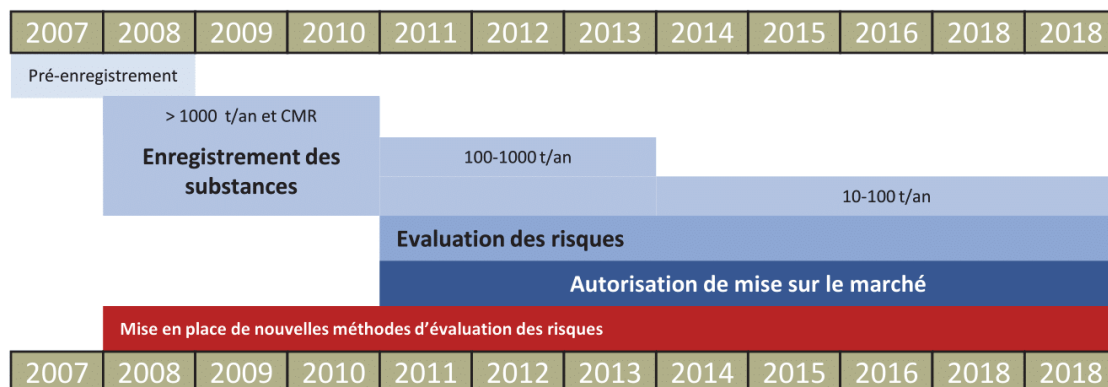


Figure 15. Planning et processus d'évaluation des risques chimiques selon REACH

D'après (Penman et al. 2015). Présentation du planning sur les dix suivant la mise en place de REACH ainsi que le protocole d'évaluation des risques défini par cette réglementation.

En association à la réglementation REACH, un nouveau système de classification, d'étiquetage et d'emballage des produits chimiques a vu le jour. Ce dernier entré en vigueur en 2009 met en œuvre les recommandations internationales du système général harmonisé (SGH), système international d'étiquetage des matières dangereuses destiné à unifier les différents systèmes nationaux. Il définit les nouvelles règles en matière de classification, d'étiquetage et d'emballages des composés chimiques en Europe. Il définit 28 classes de dangers associées à 9 pictogrammes différents.

En 2009, l'UE a adopté une nouvelle directive visant à encadrer la mise sur le marché et à améliorer l'évaluation des produits phytosanitaires. Cette directive entrée en vigueur en 2011 renforce les critères d'exclusion lors de l'évaluation des substances par l'ECHA, notamment en excluant tout produit phytosanitaire susceptible d'être un PE. Enfin cette directive ajoute une procédure d'homologation des coformulants présents dans les produits phytosanitaires (adjuvant, synergistes ou substances destinées à être utilisées dans un produit phytopharmaceutique) et encourage la substitution de substances actives dangereuses pour l'Homme et l'environnement.

En 2012, c'est la catégorie des biocides qui fit l'objet d'une nouvelle directive. Cette catégorie regroupe tous les produits utilisés pour protéger l'Homme, les animaux, le matériel ou toute autre substance contre les nuisibles et micro-organismes. La réglementation sur les produits biocides (RPB) a pour objectifs d'uniformiser le marché des phytosanitaires dans l'UE et de simplifier les procédures d'autorisation des substances actives et des biocides, ainsi que d'établir de nouveaux calendriers d'évaluation et de prise de décision. Chaque substance active autorisée l'est pour une durée limitée de 10 ans. Pour que la mise sur le marché d'une substance biocide soit acceptée, elle doit obtenir une autorisation de tous les états membres de l'UE.

Enfin, pour faire face au côté ubiquitaire des produits phytosanitaires, en 2008 la France a renforcé les mesures de protection de l'environnement et de l'Homme face à ces produits en mettant en place le plan Ecophyto I. Ce plan, élaboré dans le cadre du Grenelle de l'Environnement, vise à réduire d'ici 2018 la consommation de produits phytosanitaires de 50% par la suppression progressive des substances les plus préoccupantes (Plan Ecophyto 2008). En 2015, le bilan très contrasté de cette initiative (Hossard et al. 2017) a conduit à un report de cet objectif à 2025 dans le cadre du plan Ecophyto II (Plan Ecophyto II 2015). Ce plan prévoit, entre autres, l'interdiction de l'utilisation de produits phytosanitaires dans les espaces publics pour 2017 et l'interdiction de l'usage par des particuliers de pesticides à partir de 2019.

1.4.2. Règlementation non adaptée aux PEs?

Au début du seizième siècle, le médecin et philosophe suisse Paracelse (de son vrai nom, Theophrast Bombast von Hohenheim) énonçait la phrase suivant : « c'est la dose qui fait le poison ». Au-delà du côté historique de cette citation, c'est sur ce principe qu'est fondée la réglementation en vigueur en matière de toxicologie, y compris, a priori, en ce qui concerne les PEs. Ce principe est aujourd'hui remis en cause par la communauté scientifique qui estime que ce principe ne s'appliquerait pas toujours pour certains PEs. En effet, actuellement les autorités imposent le concept de dose journalière admissible (DJA) afin d'estimer les effets délétères des composés chimiques sur l'organisme. Ce concept, introduit en 1961 par le Pr. René Truhaut, a été adopté par défaut par plusieurs agences gouvernementales telles que l'Union Européenne, l'OMS ou encore la FDA. La DJA se définit alors comme la dose journalière admissible, exprimée en mg de substance par kilo de poids corporel, qu'un individu de 60 kg peut ingérer quotidiennement sans risque pour la santé. Le calcul de cette DJA fait appel à la notion de dose sans effet (DSE) ou NOAEL pour *No Observable Adverse Effect Levels* en anglais (Nair and Jacob 2016). Dans un premier temps, une évaluation de la toxicité de la substance est réalisée à forte dose. Si un effet toxique est constaté à forte dose, la substance est ensuite testée à des doses de plus en plus faibles jusqu'à la détermination de la LOAEL (*Lowest Observed Adverse Effect Levels* ; la dose minimale avec effet nocif observé) puis de la NOAEL, à laquelle il est estimé que la substance n'a aucun effet sur l'organisme. La DJA est alors calculée en divisant la NOAEL obtenue par un facteur d'incertitude dépendant de la substance étudiée. Ce facteur est généralement fixé à 100, mais si les données associées à la substance étudiée ne sont pas suffisantes ou que les effets causés sont irréversibles, ce facteur peut être revu à la hausse allant jusqu'à 1000. Inversement si le composé est bien étudié et que son impact sur l'Homme est bien documenté, il peut être revu à la baisse. Le facteur d'incertitude est destiné à apporter un niveau de sécurité supplémentaire en prenant en compte les variations intra- et interspécifiques, notamment lorsque les effets d'une substance sont étudiés sur un organisme différent de l'Homme. L'utilisation de la dose pour déterminer la toxicité d'une substance sous-entend que plus l'échantillon va être exposé à un produit chimique et plus l'effet toxique (si effet toxique il y a) observé sera important. Cette hypothèse est matérialisée graphiquement par une droite appelée courbe dose réponse. En toxicologie, cette courbe ne ressemble pas à une droite, mais à une sigmoïde, mettant en avant un effet seuil, souvent la mort de l'organisme, de la toxicité du composé testé (Figure 16). L'étude de certaines substances a cependant mis en évidence des formes de courbe différentes, non plus sigmoïdales, mais en forme de cloche ou de cloche inversée. Ces courbes dites non monotones sont généralement liées à des mécanismes propres à l'endocrinologie tels que la cinétique

d'interaction entre l'hormone et son récepteur, à l'action de mécanisme propre à la forte ou la faible dose, la demi-vie des hormones ou encore leurs transports (pour revue, Vandenberg et al. 2012).

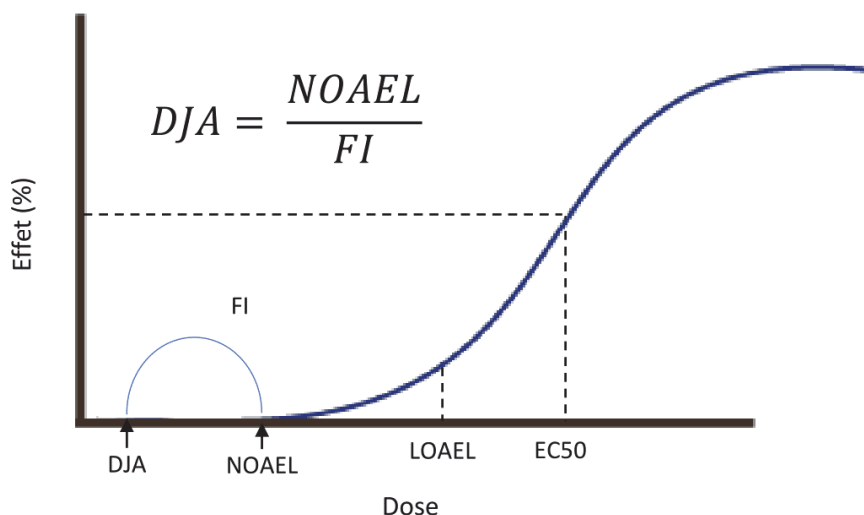


Figure 16. Représentation d'une courbe dose-réponse et calcul de la DJA

Courbe dose réponse monotone. À partir de cette courbe plusieurs indicateurs sont évalués. L'EC50 (*half maximal Effective Concentration*) est la concentration à laquelle 50% de l'effet maximal de la molécule est observé. La LOAEL (*Lowest Observed Adverse Effect Levels*) correspond à la dose minimale avec effet nocif observé. Enfin la NOAEL (*No Observable Adverse Effect Levels*) est la dose pour laquelle aucun effet nocif est observé. À partir de cette dose, la dose journalière admissible (DJA) est calculée en divisant la NOAEL par un facteur d'incertitude (FI).

Enfin, comme détaillé précédemment, un PE peut avoir un effet tissu/cellule spécifique, affectant uniquement les cellules et/ou organes sensibles à l'hormone ou la famille d'hormones cibles; Il peut également exercer des effets néfastes importants à différents stades du développement de l'organisme; il peut parfois avoir un effet retard visible uniquement des années, voire des dizaines d'années après l'exposition; il peut même dans certains cas exercer un effet transgénérationnel, c'est-à-dire avoir un effet néfaste sur l'individu exposé, mais sur sa descendance, voire la descendance de sa descendance; enfin, il peut avoir un effet cocktail, c'est-à-dire agir en synergie ou de façon antagoniste avec d'autres molécules (PEs ou non) présentes dans l'environnement. Face à ces particularités des PEs, il a été nécessaire d'adapter les outils traditionnellement utilisés en toxicologie afin de mettre en place des tests de détection qui leur sont spécifiques.

1.4.3. Outils de détection de toxicités

S'il existe des tests de détection de toxicité spécifique pour les PEs, il existe également d'autres outils utilisés par les toxicologues pour déterminer si une molécule est néfaste pour l'organisme ou non. L'identification de cette toxicité s'effectue aussi bien *in vitro*, *ex vivo*, *in vivo* qu'*in silico*. Ne pouvant être exhaustif sur l'ensemble de ces approches, celles spécifiquement dédiées aux PEs (cf. 1.4.4. Tests spécifiques de détection des perturbateurs endocriniens) et/ou relevant de la toxicologie prédictive (cf. 1.5. La toxicologie prédictive) seront plus particulièrement détaillées dans ce manuscrit.

Afin d'évaluer les effets nocifs d'un composé sur l'organisme les toxicologues utilisent classiquement la mesure de paramètres tels que les variations de poids d'un organe ou d'un tissu, la modification intra-tissulaire via des observations macro/microscopiques, l'évaluation de la mort cellulaire, le dosage d'hormones ou d'autres métabolites dans les fluides corporels. Chacune de ces observations est corrélée avec les autres données récoltées afin de permettre une interprétation globale. La présentation exhaustive de toutes les techniques usuellement utilisées en toxicologie étant en dehors de mon sujet, je me limiterai uniquement à quelques exemples.

1.4.3.1. Le test d'Ames

Le test d'Ames repose sur l'utilisation d'une souche bactérienne de *Salmonella typhimurium* mutante pour les gènes nécessaires à la synthèse d'histidine, acide aminé que ces bactéries ne peuvent donc synthétiser (His⁻). Il permet d'estimer le potentiel mutagène d'un composé chimique par réversion de ce phénotype (His⁺) (Ames et al. 1973). Cette technique simple, sensible et précise offre la possibilité d'étudier aussi bien les composés chimiques industriels, que les eaux à usage alimentaire, les gaz d'échappement, les médicaments et leurs principes actifs ainsi que les fluides biologiques. Cependant ce test reposant sur l'utilisation de bactéries le problème de transposition au modèle humain se pose. Des tests plus adaptés, dits "d'aberration chromosomique", peuvent ainsi être réalisés chez les mammifères, *in vitro* (cultures cellulaires) ou *in vivo* (lignée hématopoïétique). Une étude toxicologique incluant une exposition longue est nécessaire pour réfuter un test d'Ames positif.

1.4.3.2. Les cultures cellulaires

L'utilisation de cultures cellulaires en toxicologie repose sur la possibilité de limiter la différenciation cellulaire en conditionnant leur micro-environnement (matrice extracellulaire) et les facteurs de croissances et ainsi obtenir un type spécifique de cellules. Cet outil permet d'avoir accès facilement des cellules humaines et offre la possibilité d'étudier les mécanismes d'actions dans lesquels elles sont

impliquées. Cette méthode fait partie de la procédure standard de criblage des composés chimiques et plus particulièrement de l'évaluation des risques en pharmacologie. Une autre utilisation de ces cultures cellulaires concerne les cellules souches embryonnaires. Il s'agit d'un bon modèle de reprotoxicité puisqu'elles représentent la phase la plus précoce de l'ontogenèse. À la fin des années 90, l'ECVAM (*European center for the validation of alternative methods*) a mis au point l'*embryonic stem cell test* (Spielmann et al. 1998). Ce test utilise des cellules souches embryonnaires murines afin d'estimer le potentiel cytotoxique d'une molécule et son effet sur la différenciation cellulaire. À partir de ce test, il est possible d'estimer l'ID50, c'est-à-dire la concentration d'une substance qui inhibe de 50 % la différenciation spontanée des cellules souches embryonnaires en cellules contractiles. La mesure de ce paramètre est souvent remplacée par la quantification de l'expression de marqueurs de différenciation. La cytotoxicité, elle, est évaluée en déterminant l'IC50, à savoir la concentration d'une substance inhibant de moitié la croissance des cellules souches embryonnaires par rapport à des cellules d'une lignée adulte contrôle. Une étude ayant appliqué cette technique à 20 produits toxiques caractérisés *in vivo* a ainsi pu en classer correctement 78 % au sein de trois catégories : non toxiques, faiblement toxiques et fortement toxiques (Genschow et al. 2004). Les cultures cellulaires permettent également de reconstituer des tissus *in vitro*. Cela est déjà largement utilisé dans l'industrie pharmaceutique qui utilise des cultures de cellule de peau humaine pour prédire l'effet toxique cutané d'un produit et le franchissement de cette barrière (Robinson, Osborne, and Perkins 2000). Cependant les techniques d'évaluation basées sur la reconstitution cellulaire, la coculture ou encore la culture en trois dimensions nécessitent une grande technicité sont plus difficiles et coûteuses à mettre en place que des cultures cellulaires classiques, elles n'en restent pas moins porteuses d'avenir, représentant au mieux la situation du tissu *in vivo* ainsi que la complexité des interactions cellulaires.

Le passage de la culture cellulaire au modèle *in vivo* pose souvent problème et seulement 10% des composés testés présentent un effet similaire observé en culture et dans l'organisme, dans lequel la toxicité est plus importante que celle prédite *in vitro* (DiMasi and Grabowski 2007). Cela est dû au fait que l'effet de la molécule est testé uniquement sur un seul type de cellule, non représentatif des interactions entre les différentes cellules, mais également que celles-ci sont en culture sur une surface plane, ne prenant pas en compte la conformation spatiale tridimensionnelle des cellules (Breslin and O'Driscoll 2013). Pour pallier à ces problèmes, de nouvelles méthodes ont été mises en place afin de prendre en compte les interactions entre plusieurs types cellulaires, on parle de coculture (hépatocytes/cellules biliaires, cellules alvéolaires/macrophages, kératinocytes/fibroblastes). Mais également pour considérer la structuration dans un espace en trois dimensions : les cultures cellulaires

3D, fournissant des informations plus précises et pertinentes d'un point de vue physiologique augmentant ainsi le pouvoir prédictif de cette méthode (Birgersdotter, Sandberg, and Ernberg 2005).

Pour finir un tout nouveau domaine est en pleine émergence : les organes artificiels. Ces technologies mettent en relation dans un système de perfusion plusieurs tissus d'un même organe. Cette approche complexe, actuellement au stade d'étude fondamentale pourra dans un avenir proche fournir des informations sur la dynamique de la molécule au sein de l'organisme, informations jusque-là accessibles uniquement de façon *in vivo* (Cho and Yoon 2017).

Si la majorité de ces techniques sont usuellement utilisés en toxicologie classique pour évaluer la toxicité d'une substance chimique, il est possible également de les utiliser dans le cadre de la toxicologie prédictive afin de prédire un effet délétère.

1.4.3.3. Les cohortes

Les modèles toxicologiques ne sont pas forcément pertinents vis-à-vis de la santé humaine, car ils sont effectués sur une autre espèce et sur un seul type cellulaire et qu'ils ne prennent pas en compte la voie d'exposition aux composés chimiques, dans quelle proportion et sur quelle période.

Chez l'Homme, l'expérimentation directe des molécules étant évidemment impossible, il est difficile de relier les pathologies observées à l'exposition des perturbateurs endocriniens. Le problème soulevé par ces substances a poussé les scientifiques et les agences sanitaires à mettre en place un certain nombre d'études épidémiologiques et l'établissement de cohortes. Une cohorte est un ensemble de sujets présentant certaines caractéristiques communes, et qui sont suivis sur plusieurs années. Il existe deux types de cohortes : les cohortes prospectives et les cohortes rétrospectives. Les cohortes prospectives sont des modèles épidémiologiques d'observation dans lequel un groupe exposé et un autre non exposé sont suivis pour comparer l'incidence de la pathologie étudiée. Les cohortes rétrospectives sont des modèles d'observation dans lequel on compare le risque de contracter la maladie pour un groupe exposé et pour un autre non exposé. La cohorte PELAGIE (Perturbateurs Endocriniens : Étude Longitudinale sur les Anomalies de la Grossesse) a par exemple été mise en place par l'INSERM afin de répondre aux préoccupations de santé dues à la présence des perturbateurs endocriniens dans notre environnement quotidien. Cette cohorte, débutée en 2002, suit 3421 mères et leurs enfants en Bretagne (Béranger et al. 2017; Viel et al. 2017). Les cohortes permettent de générer une quantité importante de données, mais leur mise en place soulève néanmoins de considérables difficultés. La première est la constitution de la cohorte. La création de celle-ci se fait sur la base du volontariat et doit être composée d'un nombre d'individus suffisamment important pour permettre l'étude des différents facteurs et tout en écartant un

certain nombre d'effets confondants. Le suivi de ces cohortes peut être un facteur limitant. Durant toute la durée de l'étude (plusieurs années), les patients peuvent être amenés à déménager, ou tout simplement à abandonner l'étude. D'un point de vue expérimental, il existe aussi un décalage entre l'exposition étudiée (le plus souvent évaluée via des questionnaires ou bien mesurée et l'exposition réelle. Par ailleurs, ces études se focalisent généralement sur un voire quelques composés, alors qu'en réalité, l'organisme est exposé à un mélange de composés. Au final, et sans évidemment remettre en question la nécessité absolue de ces cohortes, les études épidémiologiques mettent en évidence des associations entre des pathologies et des composés, sans pour autant permettre l'établissement de liens de causalités.

1.4.4. Tests spécifiques de détection des perturbateurs endocriniens

Dans l'optique d'améliorer la détection et l'évaluation des PE, l'OCDE (Organisation de Coopération et de Développement Économiques ; OECD en anglais) et US EPA aux États Unis ont mis en place différents programmes. Ils ont pour objectifs de coordonner les activités, d'élaborer de nouvelles lignes directrices pour les essais et d'harmoniser les méthodes de détection et d'évaluation des risques. En 2002, l'OCDE a émis un cadre conceptuel (*Conceptual Framework*) découpé en cinq niveaux pour le test et l'évaluation des PE. Sur son site, l'OCDE précise bien que ce cadre ne constitue pas une stratégie de dépistage des PE, mais est plutôt le reflet des informations obtenues par les tests préconisés aux différents niveaux. Ce cadre a été mis à jour en 2012 (OECD 2012), amenant ainsi le nombre de tests officiels à 50, dont 12 ont été spécifiquement développés ou mis à jour pour la détection des PE et sont détaillés ci-dessous (Tableau II).

N° essai	Titre	Année
440	Essai de dépistage à court terme des propriétés oestrogéniques	2007
407	Toxicité orale à doses répétées - pendant 28 jours sur les rongeurs	2008
211	Daphnia magna, essai de reproduction	2011
441	Essai de dépistage à court terme de propriétés (anti)androgéniques	2009
229	Essai à court terme de reproduction des poissons	2009
230	Essai de 21 jours sur les poissons : dépistage à court terme de l'activité oestrogénique, et androgénique et de l'inhibition de l'aromatase	2009
231	Essai de métamorphose des amphibiens	2009
455	Essais in vitro de transactivation par transfection stable visant la détection des substances agonistes et antagonistes des récepteurs des œstrogènes	2009
233	Essai de toxicité sur le cycle de vie des chironomes dans un système eau-sédiment chargé ou eau chargée-sédiment	2010
234	Essai de développement sexuel des poissons	2011
456	Essai de stéroïdogénèse H295R	2011
457	Essai de transactivation faisant appel au récepteur des œstrogènes BG1Luc pour identifier les agonistes ou antagonistes des récepteurs des œstrogènes	2012

Tableau II. Tests officiels de l'OCDE pour la détection de perturbateurs endocriniens

1.4.4.1. Essais *in vitro* fournissant des données sur le mécanisme endocrinien perturbé

Deux essais à visée mécanistique sont recommandés par l'OCDE : les essais ER BG1Luc et les essais sur la stéroïdogénèse via la lignée cellulaire H295R, lesquels reposent principalement sur l'utilisation de lignées cellulaires.

Les **essais ER BG1Luc** ont pour objectif d'identifier des agonistes ou antagonistes des récepteurs aux œstrogènes en utilisant la lignée cellulaire BG1Luc4E2 dérivée d'adénocarcinomes ovariens d'origine humaine (Rogers and Denison 2000). Ils reposent sur la fixation d'une substance chimique étudiée sur un récepteur donné, entraînant ou non sa transactivation et la synthèse d'un produit de gène rapporteur (Sonneveld et al. 2006; Escande et al. 2006). Dans le cas de ces essais, le gène rapporteur de la luciférase est sous contrôle des récepteurs aux œstrogènes. Ils permettent de démontrer l'activité agoniste ou antagoniste d'une molécule sur ces récepteurs, en fonction de la dose et du temps.

La méthode d'essai de la **stéroïdogénèse sur la lignée cellulaire H295R** a pour objectifs de détecter les effets d'une molécule sur les composants de la voie de synthèse des stéroïdes et plus particulièrement sur la production de 17 β -œstradiol et de testostérone. La lignée cellulaire H295R a été isolée à partir d'un carcinome surrénalien d'origine humaine (Wang and Rainey 2012). Elle exprime toutes les enzymes clés de la stéroïdogénèse et produit ainsi des corticostéroïdes et des stéroïdes sexuels. Cet essai n'a cependant été validé que pour la détection de la testostérone ou du 17 β -œstradiol (Hecker et al. 2006).

1.4.4.2. Essais *in vivo* fournissant des données sur le mécanisme endocrinien perturbé

L'OCDE propose par ailleurs plusieurs tests ayant pour objectif de dépister *in vivo* l'activité oestrogénique, androgénique et l'inhibition de l'aromatase induite par une molécule, chez les mammifères, les poissons ou encore les amphibiens. Il s'agit chez les mammifères des essais utérotropiques et des essais de Hershberger, chez les poissons des essais à court terme de reproduction de 21 jours et chez les amphibiens, des essais sur la métamorphose.

Le **test utérotropique chez les rongeurs** est un test de dépistage à court terme développé dans les années 30 (Dorfman, Gallagher, and Koch 1935) et adopté par l'OCDE le 16 octobre 2007. Il se fonde sur l'augmentation du poids utérin - ou réponse utérotropique - et permet d'évaluer la capacité d'un produit chimique à induire une activité biologique analogue à celle des agonistes des œstrogènes naturels. La réponse initiale est un accroissement pondéral dû à l'absorption d'eau, qui est suivie d'une

augmentation du poids due à la croissance tissulaire (R. C. Jones and Edgren 1973). La réponse de l'utérus chez le rat et la souris est qualitativement comparable.

Le **test de Hershberger**, adopté le 7 septembre 2009, est un test de dépistage à court terme utilisant des tissus accessoires de l'appareil reproducteur mâle du rat. Il se base sur la variation de poids de 5 tissus androgéno-dépendants (prostate ventrale, vésicule séminale, muscles élévateurs de l'anus et bulbocaverneux, paire de glandes de Cowper et gland) et évalue la capacité d'une substance à mimer l'activité des agonistes ou des antagonistes d'androgènes ou d'inhibiteurs de la 5 α -réductase (Hershberger, Shipley, and Meyer 1953). Chez le rat mâle castré, ces cinq tissus répondent tous aux androgènes par une augmentation de leur poids. Au cours des années 60, ce test a permis d'évaluer plus de 700 androgènes putatifs, faisant de ce test une méthode de référence pour la détection de composés aussi bien androgéniques qu'anti-androgéniques (J Kanno et al. 2001; Jun Kanno et al. 2003a, 2003b; Owens et al. 2007; Moon et al. 2010).

L'**essai d'exposition de 21 jours** consiste à mesurer deux paramètres chez les poissons, aussi bien mâles que femelles : Le premier est le dosage de la vitellogénine (effectué chez le poisson-zèbre, le poisson-tête-de-boule et le medaka japonais), précurseur des protéines du vitellus, qui permet de détecter si le composé chimique a une action oestrogénique ou anti-oestrogénique (Sumpter and Jobling 1995). Le second paramètre est le développement de caractères sexuels secondaires chez le medaka japonais et le poisson-tête-de-boule, ces caractères étant réactifs aux taux d'androgènes présents dans l'eau de sexe opposé (Ankley et al. 2003; Seki et al. 2004). Enfin, la fécondité est également évaluée ainsi que l'adaptation du système reproducteur par l'analyse de coupes histologiques des gonades.

Enfin l'**essai de métamorphose** est un essai visant à identifier les substances capables de perturber l'axe hypothalamo-hypophyso-thyroïdien chez les amphibiens. La métamorphose des amphibiens dépendants de la thyroïde et répond aux substances actives sur cet axe. Des têtards sont exposés à au moins trois concentrations de la substance étudiée, et ce pendant 21 jours. À l'issue de ces 21 jours plusieurs paramètres dont l'histologie de la thyroïde, la longueur des pattes postérieurs et longueur museau-cloaque, sont évalués pour déterminer l'effet de la molécule sur l'axe hypothalamo-hypophyso-thyroïdien.

1.4.4.3. Essais *in vivo* fournissant des informations sur la toxicité endocrinienne

Plusieurs tests validés par l'OCDE permettent d'évaluer les effets d'une exposition chronique et de longue durée, aussi bien en termes de toxicité que d'impact sur le développement de caractères sexuels secondaires.

L'essai de toxicité orale à doses répétées pendant 28 jours sur les rongeurs a ainsi pour but d'administrer quotidiennement la substance à tester sur plusieurs groupes d'animaux afin d'identifier les organes cibles de la molécule, les doses administrables ainsi que le calcul de la NOAEL et/ou LOAEL. Chaque jour, les animaux sont inspectés afin de déceler des signes potentiels de toxicité. Cet essai permet de récolter de nombreuses données sur les tissus, et les glandes endocrines en particulier, via des coupes histologiques, mais également le poids de ces organes, la qualité spermatique ou la normalité des cycles œstraux.

Les **tests pubertaires** consistent quant à eux à traiter durant une période de 20 à 30 jours des rats mâles ou femelles afin de déterminer l'âge d'apparition des premiers signes de puberté (séparation du prépuce, ouverture vaginale). Ils permettent également de détecter les produits chimiques possédant une activité androgénique ou anti-androgénique.

1.4.4.4. Essais *in vivo* de compréhension approfondie

La combinaison des effets délétères relativement fréquents des PEs sur la reproduction d'une part, et des potentiels effets transgénérationnels de ceux-ci d'autre part, ont amené l'OCDE à considérer plusieurs tests prenant en compte ces deux aspects dans l'évaluation des PEs.

L'essai étendu de toxicité pour la reproduction sur une génération a pour objectif l'évaluation des effets pré et postnataux sur la reproduction et le développement après une exposition à une substance chimique. Cet essai est couramment couplé avec les essais utérotropiques et de Hershberger. L'examen détaillé des principaux effets sur le développement, comme la viabilité de la descendance et le développement physiologique et fonctionnel jusqu'à l'âge adulte, doit permettre de repérer les organes cibles et le mode d'action de la molécule chez les descendants.

Le **test d'étude de toxicité pour la reproduction sur deux générations** est quant à lui conçu afin de déterminer l'impact d'une molécule sur l'intégrité et la performance des systèmes reproducteurs mâle et femelle (F0), sur la croissance et le développement de la descendance (F1), mais aussi sur l'intégrité des appareils reproducteurs chez la descendance de la descendance (F2).

1.5. La toxicologie prédictive

L'un des défis en toxicologie est de pouvoir extrapoler les résultats issus des différentes phases de l'analyse du risque à partir de systèmes expérimentaux vers les populations humaines. Les modèles animaux en particulier, bien que très utilisés, présentent souvent des différences notamment en termes de clairance des substances ou d'activité enzymatique. Pour ces raisons pratiques, mais également éthiques, politiques et économiques, des efforts importants sont demandés aux laboratoires pour Remplacer ces modèles par d'autres alternatives, Réduire au minimum leurs utilisations et Raffiner les stratégies expérimentales afin de minimiser le stress et la douleur des animaux (principe des « 3R »). C'est dans ce contexte que les organismes de prévention et de gestion des risques chimiques se sont tournés vers la toxicologie computationnelle et plus précisément vers la toxicologie prédictive. Ces méthodes consistent en l'extrapolation des informations connues associées à une molécule afin de prédire l'effet de celle-ci ou d'une molécule similaire sur l'Homme et son environnement via la détermination de sa « signature » toxicologique. Cette signature peut être d'ordre divers (physiologique, moléculaire, génomique...) sur un individu ou sa descendance après exposition à un ou plusieurs facteurs (biologique, physique ou chimique) (Robert J Kavlock et al. 2007). Cependant l'utilisation de la toxicologie prédictive dans un cadre réglementaire reste très limitée et seule l'étude des produits cosmétiques en fait l'objet depuis 2009.

Depuis une trentaine d'années, la toxicologie prédictive vise à diminuer le plus possible le nombre d'expérimentations *in vivo* chez l'animale en les remplaçant par des méthodes *in vitro* ou *in silico*, afin d'évaluer les risques pour l'Homme, de façon aussi précise, si ce n'est plus. L'utilisation de la toxicologie prédictive est particulièrement recommandée dans le cadre de la législation REACH. Différents pays de l'Union Européenne ont dans ce contexte mis en place des agences afin de répondre aux besoins présents aussi bien dans l'industrie que dans la recherche dans ce domaine. Ainsi l'ECOPA (*European Consensus Platform for Alternatives*), l'agence française de sécurité sanitaire des produits de santé (Afssaps) ou encore l'Institut National de l'Environnement Industriel et des Risques (INERIS), ont vu le jour.

1.5.1. Outils de toxicologie prédictive

Ces tests se basent sur les informations existantes dans les banques de données. À ce niveau, il s'agit de faire le bilan des connaissances en toxicologie et d'orienter les futurs tests (*in vivo*, *in vitro*) en utilisant les propriétés physico-chimiques, l'appartenance à des familles chimiques ou encore les données d'écotoxicologie pour estimer le potentiel toxique de la substance étudiée. On parle, ici, de « *non-testing* »

methods », de toxicologie computationnelle ou encore de méthodes *in silico*. D'après Hartung et ses collaborateurs, la toxicologie *in silico* est définie comme « tout ce que nous pouvons faire avec un ordinateur en toxicologie, c'est-à-dire une grande partie des tests, car la plupart utilisent la planification et/ou l'analyse par ordinateur. » (Hartung and Hoffmann 2009). La toxicologie *in silico* diffère notamment de la toxicologie traditionnelle de par l'échelle d'étude : elle permet en effet d'étudier un grand nombre de produits chimiques, de couvrir de nombreuses voies métaboliques et critères d'effets (*endpoints*) et ce sur plusieurs niveaux d'organisation biologique, tout en considérant simultanément plusieurs conditions d'exposition (R. J. Kavlock et al. 2007). Ces méthodes ont pour objectif de venir en complément des tests de toxicité *in vivo* et *ex vivo* et de diminuer l'utilisation d'animaux, de réduire les coûts et la durée des expérimentations, et d'améliorer la prédiction de la toxicité des composés étudiés. Enfin ces méthodes présentent l'avantage de pouvoir anticiper la toxicité d'un composé avant même qu'il ne soit synthétisé (Madan, Bajaj, and Dureja 2013). Parmi les outils utilisés en toxicologie computationnelle se trouvent les banques de données, les outils de description moléculaire, les simulations de dynamique moléculaire, des logiciels de génération de modèles prédictifs, des logiciels/serveurs de modèles prédictifs préalablement construits et enfin des outils de visualisation (Figure 17A). Tous ces outils sont ensuite utilisés au sein des différentes méthodes de prédiction de la toxicité.

1.5.1.1. Les bases de données

Face au nombre grandissant d'expériences de toxicologie et à l'accumulation des résultats correspondants, il a fallu mettre en place des **bases de données** sur la toxicité ainsi que des règles de classification pour les nouvelles molécules étudiées. Ces outils compilent des informations sur différents types de toxicité (néphrotoxicité, hépatotoxicité, reprotoxicité...), modes d'action ou activités moléculaires, déterminés chez différents organismes (Valerio 2009). Pour les perturbateurs endocriniens, les organismes tel que l'OCDE, l'US EPA ou encore l'organisation TEDX (*The Endocrine Disruptors Exchange*) mettent par exemple à disposition des bases de données spécifiques référençant tous les PEs avérés. Les données stockées associées à l'expertise de scientifiques permettent alors de créer un modèle de prédiction de la toxicité de la molécule et de tester celle-ci en utilisant des données issues d'autres bases (Figure 17B). D'autres banques se sont spécialisées dans la description et la mise en relation entre les composés chimiques et les pathologies en se basant sur diverses informations.

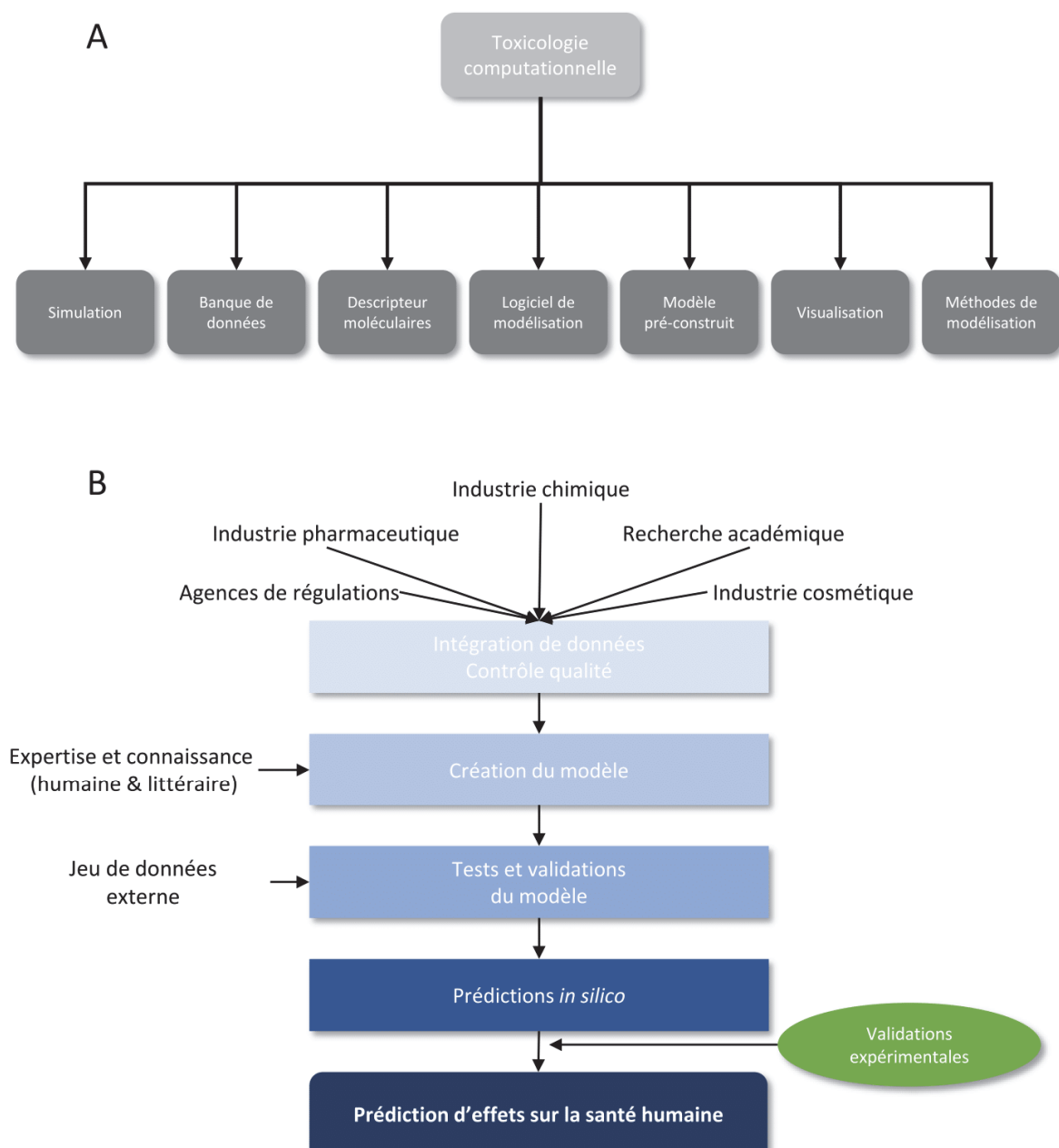


Figure 17. La toxicologie computationnelle et les bases de données

A) Les différents outils utilisés en toxicologie computationnelle. B) Utilisation des bases de données dans le cadre de la détection et de la prédiction d'effet d'une molécule.

Parmi ces dernières on peut citer la Comparative Toxicogenomics Database (CTD) (A. P. Davis et al. 2015) ou encore ChemProt (Kringelum et al. 2016). Enfin les banques de données comme PubChem (Kim et al. 2015), ChemSpyder (Williams and Tkachenko 2014) ou ChEmbl (Davies et al. 2015) ont pour objectif de concaténer toutes les informations relatives aux substances chimiques (propriétés physico-chimiques, nomenclature, description, structure...).

Ces données présentent cependant des limites. En effet, chacune des toxicités décrites dans les bases de données est obtenue à partir d'un protocole différent, utilisant des unités de mesure différentes que ce soit pour le temps ou pour la dose, ainsi qu'une terminologie propre. Bien qu'importante d'un point de vue quantitatif, la recherche d'informations dans ces bases est difficile et la comparaison entre différentes expériences l'est d'autant plus. Ceci pose deux questions sur l'utilité de ce type d'outil : où peut-on obtenir l'information sur la toxicité du composé chimique étudié, et dans quelle mesure peut-on se fier aux données ? La résolution de ces problèmes est au centre des objectifs de certaines organisations internationales telles que la FDA (Food and Drug Administration) américaine, le CDER (Center for Drug Evaluation and Research) de la FDA et la réglementation REACH dans l'Union Européenne (Mullard 2017). Toutefois, bien qu'un certain nombre d'organisations soient impliquées, l'harmonisation des informations sur la toxicité chimique entre les bases de données reste un besoin critique, mais difficile à mettre en place (Myshkin et al. 2012).

1.5.1.2. Logiciels de description moléculaire

Les **logiciels de description moléculaire** permettent de déterminer les caractéristiques physico-chimiques et pharmacologiques d'une molécule, sur la seule base de sa structure. Ce système permet de représenter la molécule sous la nomenclature de l'UICPA (Union Internationale de Chimie Pure et Appliquée), offrant une représentation lisible de sa structure et de ses constituants. Les caractéristiques telles que la charge, la solubilité ou la masse molaire sont aisément calculées à partir de la structure de la molécule, et il existe de nombreux logiciels utilisant des bases de données moléculaires permettant de définir ces descripteurs tels que Derek Nexus (Marchant, Briggs, and Long 2008), MC4PC (Contrera et al. 2007) ou encore Toxtree (Patlewicz et al. 2008).

1.5.1.3. Outils de simulation

Les **outils de simulation de dynamique moléculaire** sont des outils très importants pour la compréhension des bases structurales et fonctionnelles d'une macromolécule. Ils ont pour objectifs de répondre à des questions relatives au devenir d'une molécule dans l'organisme, par exemple

l'identification des voies métaboliques impliquées sa dégradation, la cinétique de la molécule dans l'organisme ou encore le potentiel de docking. Le nombre d'études utilisant ce type d'outils de simulation a fortement augmenté ces dernières années, notamment grâce à la disponibilité de nombreux programmes, mais également grâce à la forte évolution de la puissance informatique (Jo et al. 2017; Christen et al. 2005; Case et al. 2005).

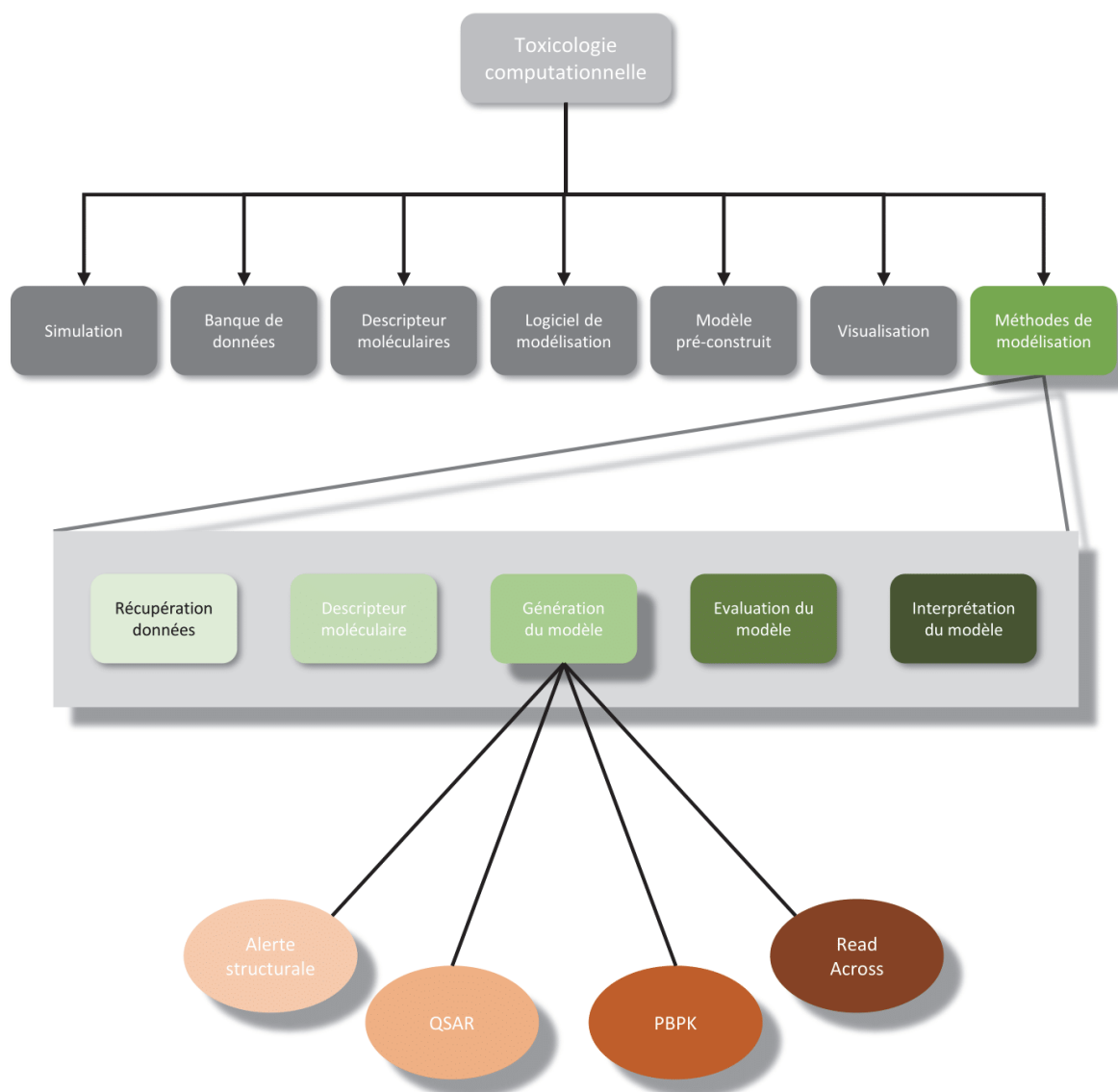


Figure 18. Génération de modèles prédictifs

Les modèles prédictifs de toxicités sont générés grâce à des descripteurs moléculaires dérivés de données expérimentales utilisant différentes méthodes : les alertes de structures, les relations structure-activité (QSAR), la pharmacocinétique (PBPK) ou les références croisées (Read Across).

1.5.2. Méthodes de prédiction

Les méthodes de prédiction se basent sur l'utilisation de l'outil informatique pour prédire l'effet d'une molécule à partir d'un modèle représentant schématiquement le ou les processus qu'elle induit ou affecte. Ces modèles sont générés grâce à des descripteurs moléculaires dérivés de données expérimentales. Différentes méthodes peuvent être utilisées pour la génération des modèles, comme les alertes de structures, les relations structure-activité (QSAR), la pharmacocinétique (PBPK) ou encore les références croisées (Read Across) (Figure 18).

1.5.2.1. Les alertes structurales

Les alertes structurales (SA), également appelées toxicophores, sont des structures chimiques identifiées comme étant associées à une toxicité (Roncaglioni et al. 2013). Un toxicophore peut être composé d'un atome ou d'une chaîne d'atomes et une molécule peut être composée d'un ou plusieurs toxicophores. Ce dernier peut ainsi voir sa toxicité démultipliée si la molécule possède d'autres structures chimiques toxiques (Lepailleur, Poezevara, and Bureau 2013). Cette méthode est usuellement utilisée sous la forme « si A est B alors T » où A est une SA, B la valeur de cette SA et T la toxicité prédite à un certain niveau. Un exemple d'utilisation de SA :

SI (*structure_chimique*) **EST** (*présent*) **ALORS** (*sensitivité_cutanée EST certain*)

La toxicité induite par une SA est établie à partir de deux modèles : un modèle basé sur la physiologie humaine (HBR) et un modèle basé sur l'induction d'informations (IBR). Les HBRs utilisent les connaissances d'experts et de la littérature. Ils sont plus précis que les IBRs mais reposent sur des connaissances pouvant parfois être incomplètes ou biaisées. La mise à jour des modèles de toxicité nécessite une analyse détaillée de la littérature et donc fastidieuse à mettre en place. Les IBRs sont obtenus à partir du recoupement d'informations présentes dans bases de données et déterminent statistiquement si une SA est toxique ou non pour un organisme. Ceci permet par exemple aux IBRs de proposer des associations entre les propriétés structurales d'un toxicophore et une toxicité, lesquelles n'auraient pas forcément été identifiées de façon manuelle.

L'utilisation des SAs est très répandue dans l'industrie pharmacologique, notamment dans la conception d'un médicament afin de déterminer comment une molécule pourrait être modifiée pour en diminuer la toxicité. Cependant les SAs présentent de nombreuses limitations. La première limitation est l'utilisation de caractéristiques qualitatives binaires (présent/absent, carcinogène/non carcinogène). De plus, elles ne fournissent pas d'informations sur les voies métaboliques impliquées dans la toxicité et ne peuvent donc être utilisées seules pour prédire une toxicité. Les SAs ne prennent pas non plus en compte

la présence ou l'absence d'autres composés chimiques pouvant augmenter ou diminuer la toxicité. Enfin, la liste des SAs ainsi que les modèles permettant leur création peuvent être incomplets et conduire à de nombreux faux négatifs (composé classé comme non toxique alors qu'il l'est toxique). Par ailleurs, ce n'est pas parce que un composé chimique ne contient pas de SA ou n'est pas associé avec une toxicité que celui-ci est non toxique. Ceci est spécialement vrai pour les HBRs pour lesquels les évidences de non-toxicité ne sont que très rarement mises en évidence dans littérature.

Les premières listes de SAs ont été publiées en 1983 par Dupuis et ses collègues et portent sur les composés présentant un potentiel d'irritation cutanée (Dupuis and Benezra 1983). En 1988 une liste de SAs pour la prédiction de composés mutagènes et carcinogènes obtenue a est publiée après que plus de 200 molécules aient été testées (Ashby and Tennant 1988). Aujourd'hui les listes de SAs les plus complètes pour la prédiction de composés carcinogènes (Benigni and Bossa 2008) sont notamment disponibles par le biais d'outils proposés par l'OCDE comme QSAR Toolbox et Toxtree (Mombelli and Devillers 2010).

1.5.2.2. Les relations structure-activité quantitatives

Les premières études sur les relations structure-activité, ou QSAR (*Quantitative structure-activity relationship*) remontent à la fin du 19e siècle. Ce n'est que dans les années 60 que cette méthode a pris de l'ampleur grâce au travail de Corwin Hansh et de ses collaborateurs qui ont proposé un modèle mathématique offrant la possibilité de corréler l'activité biologique d'une molécule et sa structure chimique. Cette technique est aujourd'hui couramment utilisée dans les premiers stades d'évaluations des risques chimiques et est encouragée par l'OCDE, l'US EPA ainsi par REACH dans le cadre de la réduction du nombre de tests animaux.

La QSAR englobe toutes les méthodes statistiques par lesquelles les activités biologiques sont reliées à des éléments de structures chimiques, des propriétés physico-chimiques ou d'autres paramètres aidant à la description de la structure. Les informations obtenues à partir des QSAR sont ensuite utilisées pour approfondir les connaissances sur les structures moléculaires et le mode d'action moléculaire du composé chimique étudié. Ces informations peuvent alors être utilisées pour prédire la toxicité d'une nouvelle molécule, mais également pour mettre en place de nouvelles structures chimiques (Gini 2016). La Figure 19 schématise l'utilisation du QSAR et son application.

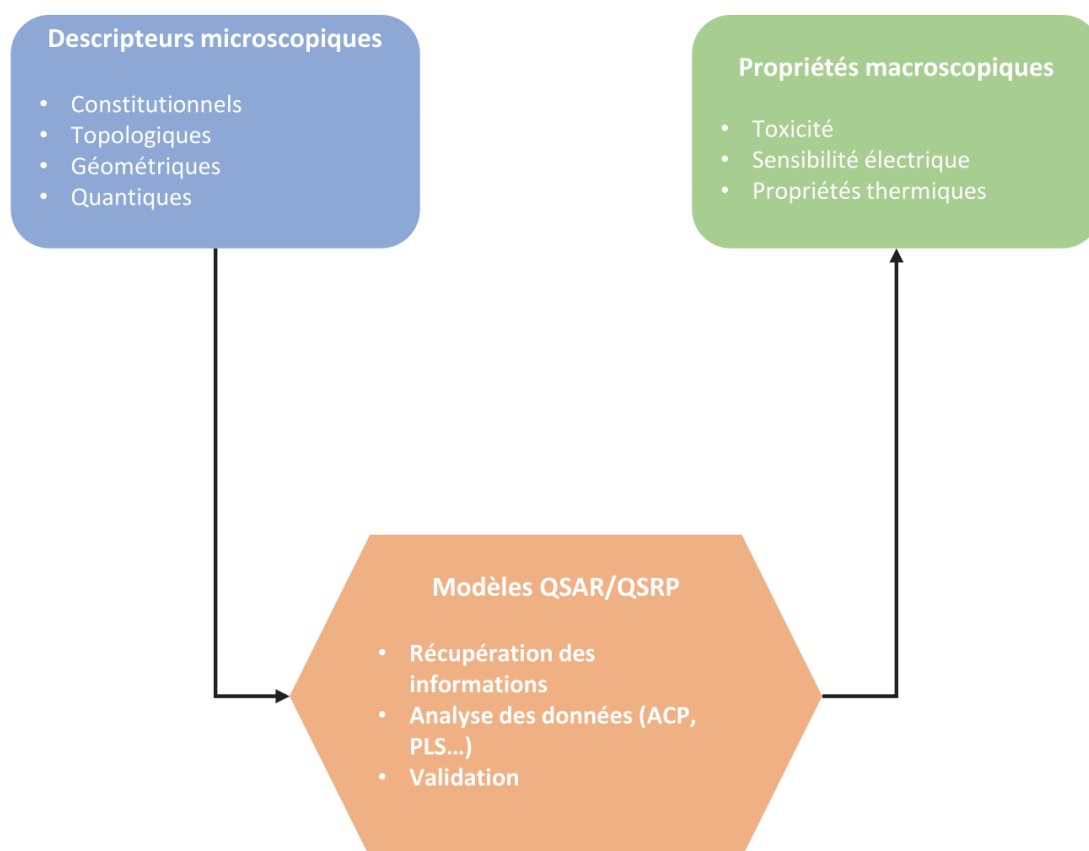


Figure 19. Création et utilisation des QSARs

Cette méthode repose sur l'utilisation de descripteurs microscopiques pour créer des modèles QSAR expliquant des propriétés macroscopiques (comme la toxicité).

En simplifiant, la QSAR utilise une fonction f qui calcul la toxicité en fonction d'un vecteur de caractéristiques chimique appelé aussi descripteur θP . Ceci se traduit par la fonction suivante :

$$T = f(\theta P)$$

Il existe deux types de QSAR. La QSAR dite 'locale' dont le modèle est généré à partir d'informations obtenues sur des produits chimiques similaires. Ces QSARs sont les plus précises, car focalisées sur une catégorie chimique précise. Cependant il n'est pas possible d'un point de vue économique, temps et puissance de calcul de développer un QSAR pour chaque type de composés chimiques. La QSAR 'globale' est quant à elle créée à partir d'un jeu de données hétérogène de composés chimiques. À la différence des QSARs locales, les QSARs globaux sont plus pratiques pour l'étude de nombreux composés, mais aussi moins précis (Valerio 2009). De plus les QSARs locales peuvent fournir des indications sur le mécanisme d'action de la molécule, ce que les QSARs globales ne peuvent faire. Il existe également des QSARs dédiées à la prédiction de la toxicologie et celle des propriétés chimiques, respectivement les QSTR (*Quantitative structure-toxicity relationship*) (L. Carlsen, Kenessov, and Batyrbekova 2008) et les QSPR (*Quantitative structure-property relationship*) (Steinmetz, Madden, and Cronin 2015).

Un modèle QSAR de type QSPR est très dépendant des données expérimentales de références. Le choix de la base de données est donc un point critique de la création du modèle. La plupart du temps ces données sont issues de la littérature, doivent être le plus homogènes possible et présenter le plus haut taux de fiabilité. Il est recommandé d'utiliser des données obtenues suivant le même protocole expérimental et qu'elles une distribution normale afin d'augmenter les performances des méthodes statistiques utilisées pour la création du modèle QSAR et de diminuer l'influence de valeurs extrêmes.

La mise en place d'un modèle QSAR nécessite au préalable la définition d'un descripteur moléculaire. Ces derniers prennent en compte les informations sur la structure et les caractéristiques physico-chimiques de la molécule étudiée. Depuis de nombreuses années, des travaux ont été menés sur ces descripteurs afin de permettre la description la plus exhaustive possible des structures moléculaires. Il existe aujourd'hui des milliers de descripteurs (Todeschini and Consonni 2008; Katritzky et al. 2002) qu'il est possible de diviser ces descripteurs en sous-classes : les descripteurs constitutionnels, les descripteurs topographiques, les descripteurs géométriques et les descripteurs quantiques.

Les **descripteurs constitutionnels** sont les plus simples pour représenter un système moléculaire. Ce dernier considère la composition chimique de la biomolécule, sans prendre en compte les conformations géométriques ou les caractéristiques électroniques de cette dernière. Parmi les descripteurs constitutionnels se trouvent le nombre (absolu ou relatif) d'atomes (C, O, H, N), le nombre

de groupes fonctionnels (OH, COOH, NO₂), le nombre de liaisons (simple, double), le nombre de cycles (aromatique ou non) ou encore la masse moléculaire. Cette simplicité en fait des descripteurs très utilisés. Ils sont employés pour obtenir des modèles de QSARs simples, mais peuvent poser problème au niveau de l'interprétation des mécanismes d'interactions mis en jeu. En effet, ce type de descripteur se retrouve très vite limité lors de l'étude d'isomères, qui ne peuvent être distingués. De plus la grande majorité des propriétés étudiées sont dépendantes de la position des substituants qui n'est, là encore, pas prise en compte par les descripteurs constitutionnels.

Les **descripteurs topologiques** ou indices topologiques sont déterminés à partir de la structure bidimensionnelle de la molécule. Cette structure représente une simple table de connectivité des atomes (vue générale sur la connexion des atomes) de la molécule et regroupe des informations sur la taille globale de la molécule, sa forme et ses potentielles ramifications. Ils s'inspirent de la théorie des graphes appliquée à la table de connectivité représentant ainsi de manière compacte les connexions interatomiques au sein de la molécule. Mathématiquement, ce descripteur considère le système moléculaire comme un graphe $G[V, R]$ où V est le nombre de sommets représentés par les atomes et R le nombre d'arrêtes correspondant aux liaisons chimiques. La distance topologique d entre deux atomes est alors définie comme le nombre minimum de liaisons reliant ces atomes. La Figure 20 présente un exemple de calcul de distance topologique.

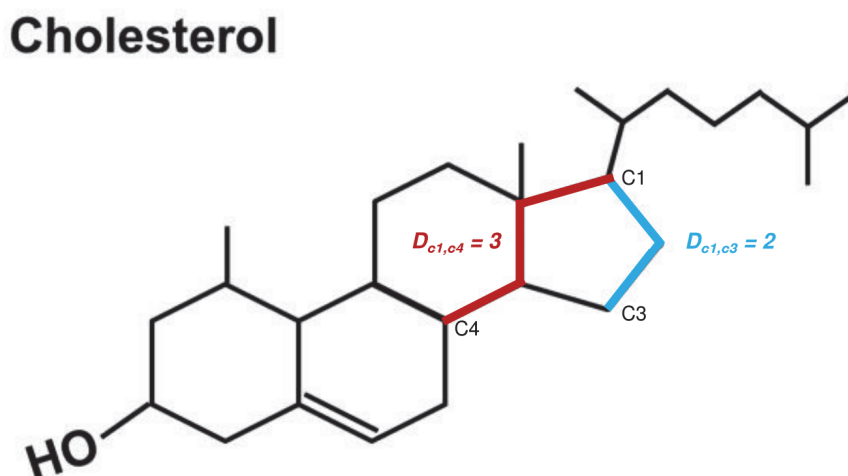


Figure 20. Calcul de distance topologique au sein de la molécule de cholestérol

La distance topologique entre deux atomes est définie comme le nombre minimum de liaisons reliant ces atomes. La distance entre les carbones 1 et 3 est égale à 2. Celle entre les carbones 1 et 4 est égale à 3.

À partir de la distance topologique, différents indices peuvent être obtenus. L'indice de Wiener permet de caractériser le volume moléculaire et le taux de ramification des molécules ; l'indice de Randic mesure l'étendue de la ramification du squelette carboné (Randic 1975) ; l'indice de Kier-Hall offre des informations sur le volume et les caractéristiques des liaisons électroniques de la molécule (Kier and Hall 1981). Les descripteurs topologiques simplifient grandement les représentations intramoléculaires en ne prenant pas en compte la distance, l'angle, la nature de la liaison ou encore la nature des atomes des molécules. Ces descripteurs sont souvent considérés comme acceptables d'un point de vue numérique, mais dans la plupart des cas, l'interprétation des équations de QSARs qui en résultent n'est pas simple, car il est difficile de relier ces descripteurs à des mécanismes biologiques.

Les **descripteurs géométriques** sont établis à partir de la position relative des atomes d'une molécule dans l'espace, nécessitant de connaître la structure tridimensionnelle de celle-ci. Cette dernière peut être obtenue expérimentalement ou par modélisation moléculaire, processus nécessitant un certain temps de calcul. Ces descripteurs comprennent le volume et la surface moléculaire ainsi que la distance et les angles entre les atomes de la molécule.

Les **descripteurs quantiques** permettent d'aller plus loin dans la description des structures moléculaires et de la quantification de caractéristiques supplémentaires. Grâce aux nouvelles avancées dans le domaine de la chimie quantique, il est possible d'avoir accès aux données énergétiques, vibrationnelles et orbitales de la molécule. À partir de telles données il est également possible de déterminer les propriétés électroniques de la molécule par exemple sa polarisabilité. Des descripteurs permettent également de quantifier les différentes interactions inter et intramoléculaires.

La mise en place de modèle QSAR de type QSTR n'est pas simple. En effet la différence d'échelle entre les données à corrélérer et les effets à prédire posent certaines difficultés, la structure étudiée se situant à l'échelle moléculaire et devant servir à prédire des effets se situant eux à l'échelle macroscopique. De plus il est nécessaire de prendre en compte l'incertitude au niveau des structures moléculaires ainsi que les incertitudes au niveau des données expérimentales. Un autre problème soulevé est celui du traitement des données à grande échelle. Il est possible d'analyser de nombreuses molécules en utilisant un grand nombre de descripteurs. Cependant aucune règle n'existe quant à la priorité à donner parmi les paramètres structuraux du jeu de données. Il existe pour cela des outils permettant de définir le moyen le plus adapté pour obtenir un modèle fiable à partir des données disponibles: les modèles linéaires par exemple la régression linéaire, la régression linéaire multiple ou encore la régression des moindres carrés pour les données continues ; les modèles non linéaires tel que les réseaux de neurones

ou encore le *support vector machine* (SVM) ; les modèles basés sur les données (*data-driven*) comme les arbres de décision, le clustering hiérarchique, la classification naïve bayésienne ou le k-mean.

Lors de la création d'un modèle QSAR de type QSTR, un grand nombre de descripteurs sont introduits, entraînant généralement une redondance de l'information et un problème de colinéarité des descripteurs. Il faut donc impérativement que les descripteurs sélectionnés soient le plus pertinent possible vis-à-vis du mécanisme étudié et autant que possible porteur d'un maximum d'informations facilement interprétables. Ceci revient à mettre en place un modèle ayant le moins de paramètres possible tout en traduisant au mieux l'information contenue dans la propriété en utilisant sur ces descripteurs des algorithmes de sélection des fonctionnalités. La sélection des fonctionnalités désigne le processus visant à réduire les entrées à traiter et à analyser, ou à identifier les caractéristiques les plus pertinentes et constitue une étape importante dans l'apprentissage automatique. Pour cela une analyse en composante principale (ACP) ou un algorithme génétique peuvent être utilisés. Enfin s'il existe peu de descripteurs pour le modèle, il est possible d'extraire les descripteurs les plus significatifs par le biais d'une analyse bidimensionnelle représentant la dispersion de chaque descripteur en fonction de l'activité biologique.

Une fois la relation propriété macroscopique-descripteurs microscopiques mise en place il est nécessaire de valider le modèle créé. Pour cela différents tests statistiques sont effectués ainsi qu'une validation interne et externe.

Pour déterminer la qualité du modèle, le coefficient de corrélation R^2 est utilisé pour estimer la part de la variance expliquée par le modèle. Plus cette valeur se rapproche de 1 et plus les valeurs prédites et observées sont corrélées. Pour évaluer la significativité statistique du modèle, l'indice de Fisher est utilisé. Enfin, la pertinence des descripteurs est estimée la plupart du temps par un test de Student.

Les validations internes reviennent à déterminer la stabilité du modèle prédictif. Pour cela, l'influence de chaque échantillon sur le modèle est estimée par l'utilisation de la technique de validation croisée. Un nombre aléatoire n de molécules est extrait du jeu initial de données à m molécules. Puis un nouveau modèle est créé à partir des $m-n$ molécules restantes en utilisant les descripteurs préalablement choisis. Le nouveau modèle obtenu est alors utilisé pour prédire les effets des n molécules soustraites. Ce processus est ensuite réitéré pour retirer et prédire les valeurs de toutes les molécules du jeu de données. D'une manière générale, les validations internes permettent l'évaluation de la robustesse du modèle, mais ne permettent pas de démontrer un pouvoir prédictif (Tropsha, Gramatica, and Gombar 2003).

Pour estimer le pouvoir prédictif du modèle QSAR obtenu, l'utilisation d'un jeu de données de validation externe est nécessaire. Ce jeu de données ne doit pas avoir été utilisé au préalable dans

l'établissement du modèle QSAR et doit être suffisamment large pour pouvoir être subdivisé en deux groupes : un groupe d'entraînement et un groupe de validation. La répartition des deux groupes est le plus souvent aléatoire. Le groupe d'entraînement va servir à entraîner le modèle de QSAR et le groupe de validation va permettre de caractériser le pouvoir prédictif du modèle.

Il existe de nombreux logiciels permettant de construire ou offrant des modèles de QSAR déjà construits. Parmi eux on peut citer QSAR Toolbox (Mombelli and Devillers 2010), TopKat (Cariello et al. 2002), DEREK Nexus (Marchant, Briggs, and Long 2008), Hazard Expert, VEGA (Pizzo et al. 2013) et METEOR.

Les résultats de QSARs ont l'avantage d'être facilement interprétables si les descripteurs sont correctement choisis. Ils peuvent modéliser les paramètres de toxicité et les descriptifs moléculaires de produits chimiques toxiques et non toxiques. L'utilisation de différents types de descripteurs permet de modéliser des activités biologiques complexes. Toutefois, il est nécessaire de disposer d'un grand nombre d'informations concernant les produits chimiques dans le développement du modèle QSAR afin qu'il puisse statistiquement significatif (Deeb and Goodarzi 2012). Il est également nécessaire de déterminer des critères de sélections des descripteurs les plus importants tout en limitant leur nombre afin de ne pas trop complexifier le modèle. Si les données utilisées pour le modèle QSAR sont uniquement structurales et non biologiques, la méthode de QSAR ne peut alors être utilisée pour l'extrapolation d'effets entre les espèces ou les voies d'expositions. Enfin les QSARs ne prennent pas en compte la dose, le temps d'exposition et l'existence de produits de métabolisation.

1.5.2.3. Les modèles pharmacocinétiques physiologiques

Les modèles pharmacocinétiques (PK) étudient la concentration d'un composé chimique dans les tissus d'un organisme en fonction du temps, estiment la quantité de substance chimique dans les différentes parties du corps et quantifient les processus d'Absorption, Distribution, Métabolisme, et Élimination (ADME). Ces modèles peuvent être empiriques ou semi-mécanistiques, compartimentés ou semi-physiologiques. Par définition un compartiment est l'ensemble ou une partie d'un organisme dans lequel la concentration (substance endogène ou exogène) est uniforme. Les modèles compartimentés représentent chaque compartiment étudié par une équation différentielle

Si les modèles pharmacocinétiques ne permettent pas de répondre de façon précise à des questions telles que « comment prédire les profils d'exposition probables dans les tissus et organes cibles ? » ou « Comment prédire les événements chez l'homme à partir de données animales ? », d'autres méthodes le peuvent. Pour pallier à ce problème, des modèles pharmacocinétiques physiologiques, ou

Les modèles PBPK sont donc des modèles mathématiques utilisés pour extrapoler et évaluer les expositions et la réponse d'un organisme en fonction de la dose d'exposition, en simulant un profil de pharmacocinétique sous de nombreuses variables physiologiques (Theil et al. 2003). Pour créer un modèle PBPK, il est nécessaire dans un premier temps de définir les paramètres à utiliser et qui sont le plus souvent dépendants de la substance testée, par exemple le coefficient de distribution tissu-plasma, la perméabilité tissulaire ou l'activité enzymatique. Certains paramètres sont en revanche indépendants, comme le flux sanguin, le volume tissulaire ou la composition tissulaire. Lors de la modélisation PBPK un descripteur mathématique de l'absorption et de la biodisponibilité d'une molécule basé sur les interrelations quantitatives avec les déterminants biologiques est développé. Il s'agit d'une description de 1) l'interaction de la molécule avec les différents compartiments tissulaires (mesure de solubilité d'un composant dans des compartiments différents) ; 2) le taux de réactions biologiques dû aux enzymes et aux transporteurs (fonctionnalité de l'organe); et 3) les caractéristiques physiologiques propres de l'organisme étudié (composition tissulaire, flux sanguin) (Rowland 2013).

Il est possible d'identifier quatre étapes dans le développement d'un modèle PBPK : 1) la représentation du modèle par une description mathématique des compartiments d'intérêts ainsi que les voies métaboliques impliquant la molécule étudiée ; 2) le paramétrage du modèle par la mise en place de paramètres mécanistiques tel que les paramètres physiologiques ou physico-chimiques ; 3) la simulation du modèle consistant à prédire l'absorption et la biodisponibilité de la molécule dans une situation définie par la résolution de l'ensemble des équations décrivant le fonctionnement des compartiments et 4) la validation du modèle par la comparaison des résultats avec des données expérimentales et une analyse d'incertitude (Haddad, Pelekis, and Krishnan 1996).

La création du modèle se fait en deux étapes. La première consiste à sélectionner les organes et/ou tissus d'intérêt et pertinents pour le modèle. Cette sélection peut être complétée par l'ajout de compartiments sanguins (veines, artères), d'organes d'excrétion (foie, reins) ou de tissu adipeux, mais également de compartiments molécule-dépendants tels que le site d'administration (intestin, poumon, peau) ou le(s) site(s) d'action (cerveau, cœur). La seconde étape est la modélisation mathématique des compartiments (M. D. Thompson and Beard 2011). Si par ailleurs le modèle PBPK étudie une molécule et ses métabolites, il est nécessaire de mettre en place des sous-modèles PBPK pour chaque métabolite et de lier les modèles entre eux par le compartiment où a lieu la métabolisation. Ce compartiment devient le processus d'entrée du sous-modèle PBPK correspondant au métabolite (Willemain et al. 2016).

Il est possible d'individualiser chaque compartiment en sous-compartiments : une section vasculaire par laquelle l'organe/tissu peut être perfusé, un espace interstitiel ou extracellulaire et un

espace intracellulaire. En utilisant cette segmentation, l'analyte arrive au tissu par la section vasculaire puis diffuse dans l'espace interstitiel avant de passer la barrière membranaire et entrer dans l'espace intracellulaire. S'il est possible de modéliser mathématiquement ces trois sous-compartiments, dans la majorité des cas ils se retrouveront agrégés pour faciliter la modélisation, car il est difficile d'avoir des informations expérimentales sur ces sous-compartiments. Lorsque le passage de l'analyte entre l'espace vasculaire et l'espace intracellulaire est rapide par rapport à l'arrivée de l'analyte par la voie sanguine, les trois sous-compartiments se retrouvent à l'équilibre. Il est alors possible de représenter ces sous-compartiments comme un seul et même compartiment. La concentration globale au sein de l'organe/tissu est représentée par l'équation simplifiée suivante :

$$\frac{dc}{dt} = \frac{Q}{V} \left(c_{in} - \frac{c}{\lambda} \right)$$

où c est la concentration de la molécule dans le compartiment, c_{in} la concentration artérielle ou d'entrée de la molécule dans le compartiment Q est le flux, V le volume de l'organe ou du tissu et λ le coefficient de partage du compartiment. Ce modèle est principalement utilisé pour l'étude d'analyte de faible poids moléculaire, liposoluble ou encore faiblement ionisé. La simplicité de ce modèle en fait également un modèle décrivant le moins bien les propriétés biophysiques des compartiments. On préférera donc à ce modèle le modèle dit à deux sous-compartiments. Ce modèle est destiné à être utilisé pour des compartiments où le passage de l'analyte de l'espace extracellulaire à l'espace intracellulaire a besoin d'être modélisé. Les tissus possédant une barrière de perméabilité sont donc modélisés avec une équation limitée à la perméabilité qui nécessite la considération de deux sous-compartiments tissulaires et est exprimée de la manière suivante :

$$\frac{dc_1}{dt} = \frac{Q}{V_1} (c_{in} - c_1) - \frac{PS}{V_1} \left(c_1 - \frac{c_2}{\lambda} \right)$$

et

$$\frac{dc_2}{dt} = + \frac{PS}{V_2} \left(c_1 - \frac{c_2}{\lambda} \right)$$

où c_1 est la concentration de la molécule dans le sous-compartiment 1, c_2 la concentration de la molécule dans le sous-compartiment 2, V_1 le volume du compartiment 1, V_2 le volume du compartiment 2 et PS le produit de surface de perméabilité. Ces équations sont une extension de l'équation du modèle à débit/perfusion limité avec la prise en compte de la perméabilité des compartiments. Les deux modèles sont représentés dans la Figure 22.

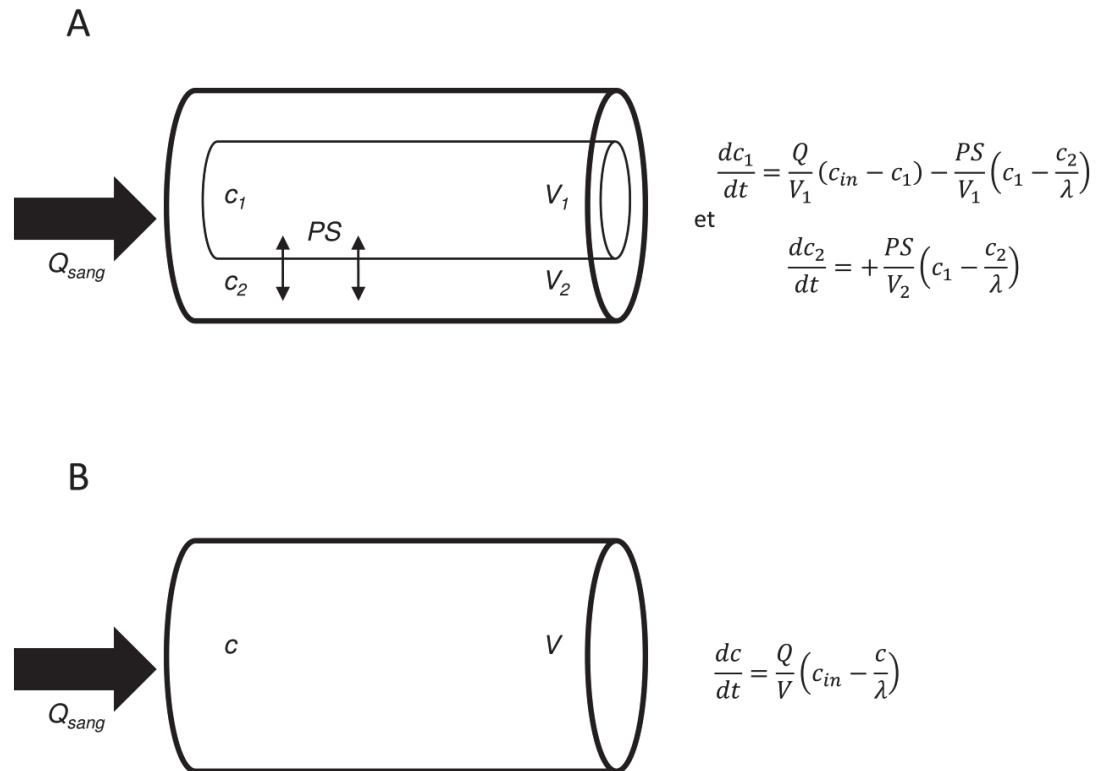


Figure 22. Représentation des modèles PBPK les plus utilisés

Représentation des deux modèles PBPK les plus utilisés. **A) Modèle à sous-compartiments.** Ce modèle est utilisé pour modéliser les interactions entre deux compartiments cellulaires par perméabilité. Il dépend de c_{in} la concentration artérielle ou d'entrée de la molécule dans le compartiment, du flux d'entrée Q , de la concentration de la molécule dans le sous-compartiment 1 noté c_1 , de la concentration de la molécule dans le sous-compartiment 2 noté c_2 , du volume du compartiment 1 V_1 , du volume du compartiment 2 V_2 , de PS le produit de surface de perméabilité et de λ le coefficient de partage du compartiment. **B) Modèle à un compartiment.** Représentation simplifiée d'un organe/tissu dépendant de c la concentration de la molécule dans le compartiment, de c_{in} la concentration artérielle ou d'entrée de la molécule dans le compartiment, de Q le flux sanguin, de V le volume de l'organe ou du tissu et de λ le coefficient de partage du compartiment.

Pour certains organes comme le foie ou le rein, il faut aussi tenir compte du processus d'élimination de l'analyte. De ce fait l'équation de base du modèle est modifiée par l'ajout d'une variable représentant le processus métabolique. Par exemple, pour simuler l'élimination d'un analyte dans le foie, l'équation mathématique est la suivante :

$$V_{FO} \frac{dc_{FO}}{dt} = Q_{FO} \left(C_{in} - \frac{C_{FO}}{\lambda} \right) - Cl'_{FO} C'_{FO}$$

où V_{FO} est le volume du foie, Q_{FO} le débit sanguin hépatique, C_{FO} la concentration totale de l'analyte dans le foie, C_{in} la concentration totale d'analyte à l'entrée du foie, λ le coefficient de partage du foie, Cl'_{FO} la clairance intrinsèque libre de l'analyte et C'_{FO} la concentration libre de l'analyte dans le foie. Enfin il est également possible de modéliser le processus d'absorption. Pour cela des logiciels comme GastroPlus sont utilisés (Bhattachar et al. 2011).

Après avoir établi le modèle et avoir modélisé mathématiquement tous les compartiments, il convient de résoudre les équations précédentes. Celles-ci requièrent de connaître trois types de paramètres : les volumes tissulaires et débits sanguins locaux (paramètres physiologiques et anatomiques), les coefficients de partage, la perméabilité cellulaire et les constantes de liaison (paramètres thermodynamiques) et les paramètres ADME. Il est malheureusement impossible d'avoir toutes les informations pour un organisme. Ces paramètres sont fixés par rapport aux informations fournies par la littérature, comme le poids moyen des organes ou les débits sanguins. Si l'information est manquante pour une espèce, on peut utiliser une approche allométrique pour fixer le poids d'un organe :

$$PE_1 = PE_2 \left(\frac{P_1}{P_2} \right)^b$$

où PE_1 est le paramètre étudié chez l'espèce 1, PE_2 est le paramètre étudié chez l'espèce 2, P_1 est le poids de l'espèce 1, P_2 est le poids de l'espèce 2 et b le coefficient d'allométrie.

Le coefficient de partage λ est dépendant de chaque analyte et doit être estimé de manière expérimentale. Cette détermination se fait *in vivo* ou *in vitro*. *In vivo* il est facile d'estimer les concentrations au sein des tissus/organes, mais le calcul de la concentration d'analyte en sortie de tissus est plus difficile, voire impossible pour certain cas. Dans ce cas la concentration dans le sang veineux de l'analyte est utilisée. Ceci comporte un biais puisque la concentration veineuse n'est pas forcément égale à celle en sortie de tissu. L'estimation du processus d'élimination est faite expérimentalement par des mesures de clairances hépatiques et rénales. Les clairances hépatiques peuvent également être estimées *in vitro* via la mesure de la clairance intrinsèque.

La disponibilité de logiciels libres ou commerciaux capables d'effectuer à la fois la simulation et l'optimisation des paramètres et, pour quelques-uns d'entre eux, des prévisions *in silico* principalement pour les processus d'absorption et de distribution offre un environnement simple pour la simulation et la validation du modèle PBPK développé. Les approches *in silico* pour la prédiction de l'excrétion reposent principalement sur des approches allométriques à travers les espèces animales. Les logiciels conçus spécialement pour la modélisation PBPK comprennent GastroPlus (<http://www.simulations-plus.com>) et PK-Sim (<https://github.com/Open-Systems-Pharmacology/PK-Sim>). D'autres logiciels informatiques de plus haut niveau sont disponibles pour les scientifiques, mais ils ont besoin de compétences en programmation : Matlab- (<http://www.mathworks.com>) ou encore R (R core Team 2013).

Les PBPK sont aujourd'hui utilisés afin de mieux comprendre la physiologie et la biodisponibilité/biodistribution d'un xénobiotique dans l'organisme. Ces modèles sont notamment très utilisés en cancérologie, par exemple pour déterminer le passage d'un anticancéreux à travers la paroi testiculaire. Cet organe représente en effet un compartiment dont la diffusion membranaire est limitée. À ce titre, il n'y aura pas de parallélisme entre les concentrations plasmatiques et testiculaires. Le but général est de maximiser la concentration de l'agent au site d'action tout en minimisant l'exposition systémique ; c'est ainsi que l'on peut modéliser l'intérêt de l'administration locale d'un médicament. Enfin les PBPK sont également très utilisés en toxicologie prédictive. En effet, ils permettent trois types d'extrapolation : l'extrapolation de l'espèce expérimentale à l'Homme, l'extrapolation d'une faible à une forte dose et l'extrapolation d'une voie d'administration à une autre.

1.5.2.4. Les Read Across / Références croisées

Le Read Across (RA) est une méthode de prédiction de la toxicité d'une molécule inconnue basée sur les connaissances relatives de molécules appartenant à la même catégorie chimique, on parle ici d'analogues chimiques (van Leeuwen et al. 2009). Si cette méthode n'est pas nouvelle, elle a été remise sur le devant de la scène depuis quelques années au vu des pressions mises sur les scientifiques pour le développement de méthodes alternatives dans le domaine de l'évaluation des risques et dans l'optique de diminuer le nombre de tests animaux. De plus l'évolution des technologies dites -omiques comme la transcriptomique, protéomique ou encore métabolomique a drastiquement changé le volume de données, accessibles la plupart du temps via des dépôts publics comme GEO (Barrett et al. 2013) ou ArrayExpress (Kolesnikov et al. 2015). À l'heure actuelle, plusieurs organismes nationaux portent un grand intérêt sur l'établissement d'une réglementation afin d'établir les meilleures pratiques et d'évaluer les résultats de RA pour permettre l'aide aux décisions réglementaires. Cette méthode a démontré via l'outil de l'OCDE QSAR Toolbox, que les RAs sont capables de prédire la toxicité d'une molécule

(Enoch, Przybylak, and Cronin 2013). Toutefois, il est en revanche plus difficile de montrer qu'une substance n'est pas potentiellement dangereuse.

Le développement de méthodes de RA se fait de deux façons : une approche analogique (*one-to-one*), utilisant un ou peu d'analogues, et très sensible aux valeurs aberrantes dues au faible nombre d'analogues utilisés, et une approche de catégorie (*many-to-one*), qui utilise de nombreux analogues. L'utilisation de nombreux analogues pour l'approche de catégorie est utile pour détecter les tendances dans une catégorie chimique et peut accroître la confiance dans les prédictions de toxicité (Vink et al. 2010). Cette méthode exige cependant de définir les limites de la catégorie permettant de déterminer si un produit chimique appartient à cette catégorie ou non et de mettre en œuvre une combinaison de prédictions (moyenne, linéaire, médiane) pour les analogues présentant des profils de toxicité contradictoires (Dimitrov and Mekenyan 2010). Enfin, la prédiction d'un effet d'une molécule est meilleure par interpolation (les informations sources entourent la molécule cible) que par extrapolation (la molécule cible est associée par supposition à partir des informations sources).

L'effet mesuré par les RAs peut être aussi bien qualitatif (présence oui ou non d'une toxicité), que quantitatif, comme la mesure d'une toxicité (puissance). L'identification des similarités entre des composés chimiques est effectuée en deux étapes : la représentation des molécules étudiées en tant que vecteurs caractéristiques de leurs propriétés chimiques ; puis le calcul de la similitude de ces produits chimiques. La première étape est obtenue en utilisant soit une « empreinte binaire », soit une « empreinte holographique ». Une empreinte binaire est la description binaire d'une caractéristique représentant la présence (1) ou l'absence (0) de cette dernière (présence d'un groupement méthyl, par exemple). L'empreinte holographique, quant à elle, utilise la fréquence d'apparition d'une caractéristique (nombre de groupements méthyl). D'autres empreintes peuvent être utilisées, par exemple des propriétés chimiques continues telles que le point de fusion ou des données génomiques comme l'expression différentielle de gènes. Il est également possible de remplacer les vecteurs caractéristiques par une hiérarchisation en catégories et sous-catégories de ces caractéristiques. Pour chaque niveau de hiérarchie, une catégorie est formée, basée sur l'empreinte des molécules permettant d'étudier la signification des propriétés et simplifiant l'interprétation du modèle. La similarité statistique de deux produits chimiques peut, ensuite, être calculée à l'aide de différents types de distances, telles que les distances de Hamming (Willett 2006), euclidienne (Andersson, Fick, and Rännar 2011), Cosine ou Mahalanobis et Tanimoto (Bajusz, Rácz, and Héberger 2015; Willett, And, and Downs 1998).

La méthode de RA possède de nombreux avantages. Le RA est transparent, c'est-à-dire qu'elle utilise des informations ou avis d'experts déjà publiés et accessibles par tous et non pas un jeu de données

privé que personne ne pourrait vérifier. Elle est également facile à interpréter et à implémenter, notamment pour les empreintes binaires. Enfin cette méthode permet de modéliser aussi bien des données qualitatives que quantitatives et utilise un grand nombre de descripteurs et d'expressions de similarité. Outre ces avantages, le RA possède aussi des limitations. La mesure de similarité entre deux composés statistiquement proches ne fournit pas d'informations biologiques sur la toxicité de la molécule. L'utilisation de mesures complexes de similarité peut rendre difficile l'interprétation du modèle par des interpolations ou extrapolations qui ne sont pas forcément significatives d'un point de vue biologique. Une autre limitation des RA est le volume de données utilisées, ce dernier étant très petit comparé à d'autres techniques comme les QSARs puisqu'il n'existe généralement pas d'analogue pour un composé donné. Enfin, cette méthode reste très dépendante des analogues chimiques utilisés, des mesures de similarité utilisées, des propriétés chimiques étudiées ainsi que des limites de classes définies. Il n'est pas rare de retrouver des résultats très différents en faisant varier ces paramètres qui restent très subjectifs et dépendants des toxicités étudiées, nécessitant la plupart du temps l'opinion d'experts pour juger des résultats obtenus. Cette approche peut être inapplicable ou inexacte si les analogues ont des profils de toxicités contradictoires ou si le nombre de produits chimiques analogues est insuffisant. Dans ces cas, l'approche QSAR peut être utilisée. Enfin, un domaine en toxicologie prédictive utilise les RAs, il s'agit de la toxicogénomique prédictive.

1.5.2.5. Toxicogénomique prédictive

Les techniques globales dites "omiques" (telles que la transcript-, proté-, métabol-, ... omique) sont de plus en plus utilisées pour caractériser les différents états des systèmes biologiques. Elles permettent de mettre au point des procédures décisionnelles globales, rapides et de faible coût relatif pour l'évaluation des risques. À l'interface de plusieurs disciplines scientifiques, ce type d'approches fait appel à des compétences variées, telles que la génomique, les statistiques, l'informatique, et bien évidemment, la nécessaire contribution des spécialistes du domaine d'application, la toxicologie. Les techniques "omiques" ont été largement et rapidement employées dans le cadre de l'évaluation du risque toxicologique à différents niveaux : génomique (étude du génome), transcriptomique (étude de l'expression des gènes), protéomique (étude quantitative et qualitative des protéines) ou encore la métabolomique (étude quantitative et qualitative des métabolites).

L'émergence de ces technologies couplées à l'expertise des toxicologues a donné naissance à la toxicogénomique. La toxicogénomique a pour objectif d'étudier l'activité du génome (gènes, protéines) après exposition à un xénobiotique. En règle générale, la toxicogénomique fait référence aux études transcriptomiques visant à caractériser l'impact d'un composé chimique sur l'expression des gènes. Ce

domaine s'est développé principalement grâce aux puces à ADN qui permettent l'analyse de l'expression des gènes de façon simultanée et ce dans différentes conditions (T. Zhou et al. 2009). Brièvement, elles permettent de quantifier les produits d'ARN produits lors de l'expression de l'ADN. La technologie de puce à ADN utilise le principe d'hybridation entre des sondes nucléotidiques (fragments d'ADN simple brin) spécifiques aux gènes présents dans le génome cible. Pour chaque hybridation entre la sonde et l'ADN complémentaire un signal est détecté et quantifié par fluorescence. Aujourd'hui la toxicogénomique s'appuie également sur les techniques de séquençage à haut débit comme le RNA-seq permettant une analyse plus profonde de l'impact génomique d'un composé. Grâce à cela, il est possible pour les toxicologues d'étudier l'effet global au niveau moléculaire d'une substance chimique et de comprendre les voies métaboliques perturbées par l'exposition à un composé.

La toxicogénomique est utilisée principalement dans deux domaines : la toxicologie mécanistique et la toxicologie prédictive. La toxicogénomique mécanistique est l'étude et l'évaluation des réponses biologiques et biochimiques face à un cas de toxicité. Cela permet d'amasser de nombreuses informations nécessaires pour l'évaluation des risques d'exposition à ces xénobiotiques dont les effets délétères se traduisent de manière très diversifiée (Hamadeh, Bushel, Jayadev, Martin, et al. 2002). L'exemple le plus connu est celui des effets génotoxiques accompagnant l'exposition à certains de ces polluants chimiques ou physiques, conduisant à l'apparition de mutations (Li et al. 2015). Cet effet pouvant être la conséquence directe de l'exposition à un agent chimique, mais également à son ou ses métabolites. Par ailleurs l'utilisation de la toxicogénomique mécanistique offre de nombreux atouts pour la génération et le test d'hypothèses de toxicité (W. Zhang et al. 2002) ou l'identification de la modification de voies métaboliques (Ghosh et al. 2015).

D'autre part la toxicogénomique a pour objectif d'utiliser les signatures géniques mises en évidence lors d'analyses mécanistiques afin de prédire la toxicité d'une nouvelle molécule. Cette prédiction repose sur l'hypothèse que deux composés chimiques présentant une signature toxicogénomique proche ont un mécanisme d'action comparable et une toxicité proche. Dans ce cadre nous définissons par signature toxicogénomique l'ensemble des gènes sur- ou sous-exprimés après exposition à une substance chimique. Cette hypothèse a notamment été émise chez le rat au début des années 2000 par de nombreux scientifiques (Bulera et al. 2001; Hamadeh, Bushel, Jayadev, DiSorbo, et al. 2002). En particulier, le travail de Steiner (Steiner et al. 2004) et ses collègues a consisté à démontrer cette hypothèse. Pour cela, ils ont exposé des rats à 28 composés hépatotoxiques et 3 composés non reconnus en tant qu'hépatotoxiques, à différentes doses et temps d'exposition. Des mesures plasmatiques d'indicateurs de toxicité ont été réalisées et, à l'issue du traitement, le foie des rats a été prélevé et soumis à une analyse histopathologique ainsi qu'une analyse transcriptomique. À partir des dosages et des

données histopathologiques, ces composés ont pu être classifiés en fonction du phénotype qu'ils induisaient. En parallèle, une méthode de classification a permis de grouper les profils transcriptomiques et de les associer à l'un des phénotypes. Enfin, une technique de prédiction supervisée de type *support vector machine* (SVM) est utilisée pour classer l'ensemble des composés en fonction de leurs signatures toxicogénomiques. Le succès de cette approche pour correctement classifier les composés connus, mais surtout pour prédire la toxicité des composés non connus a ainsi démontré le caractère prédictif des profils transcriptomiques en toxicologie. L'utilisation de ces nouvelles techniques pour l'évaluation des risques chimiques est encouragée par de nombreux programmes comme le *Human Toxosome Project* (Bouhifd et al. 2015). Cette initiative internationale lancée en 2015 a pour objectif de mettre en place un protocole d'analyse basé sur les technologies -omique (transcriptomique, protéomique et métabolomique) afin d'explorer les effets de substances chimiques sur des cultures cellulaires.

2. Objectifs de ma thèse

Mon projet de thèse s'inscrit dans la continuité des méthodes de toxicogénomique précédemment décrites pour l'évaluation des risques, et plus particulièrement celle des risques chimiques.

L'Institut de Recherche en Santé, Environnement et Travail (IRSET) dans lequel j'ai réalisé ma thèse a notamment pour mission d'étudier les facteurs environnementaux influençant la santé humaine. Au sein de ce dispositif, mon équipe d'accueil étudie les conséquences des expositions virales et chimiques sur la fertilité et sur le tractus urogénital en général. Elle a pour objectifs de répondre à des préoccupations majeures telles que la dissémination de virus par voie sexuelle, les conséquences des infections du tractus urogénital sur la fertilité et le développement de cancers, l'augmentation des anomalies du tractus urogénital masculin (cancer testiculaire, hypospadias et cryptorchidie) pouvant conduire à l'infertilité ou encore l'effet potentiellement délétère des traitements anticancéreux sur la fertilité. Pour cela, les recherches sont orientées selon 4 axes visant à étudier 1) l'exposition du tractus urogénital aux virus ; 2) la physiologie des gonades fœtales et adultes ; et l'impact sur la santé reproductive des 3) produits médicamenteux et des 4) contaminants environnementaux. C'est au sein de ces deux derniers axes que s'inscrit mon projet.

Mon travail de thèse s'appuie en grande partie sur l'hypothèse initiale de Steiner et al. (Steiner et al. 2004), à savoir la possibilité de prédire les effets toxiques d'un composé chimique sur la base de sa signature toxicogénomique, c'est-à-dire sur la base de l'ensemble des gènes dérégulés par ce composé. Toutefois, cette hypothèse a été testée dans des conditions optimales, voire biaisées : les effets hépatotoxiques des composés ont en effet été prédits au regard des composés clairement identifiés comme tels. De plus, cette prédiction a été réalisée à partir de données de transcriptomique obtenues sur l'organe cible de la toxicité étudiée, le foie. Or, dans une perspective d'aide à la décision et dans un contexte de diminution de l'expérimentation animale (principe des 3R), le caractère prédictif des signatures toxicogénomiques doit éviter l'analyse spécifique d'un organe ou tissu pour chacune des toxicités à évaluer. De plus, l'approche prédictive doit se faire sans *a priori* sur la(les) toxicité(s) des composés. Enfin, une signature obtenue chez une espèce donnée doit idéalement permettre d'extrapoler quant à la toxicité de la molécule à une autre espèce. Autrement dit, une signature toxicogénomique prédictive efficace doit être trans-organe et trans-espèce mais aussi pluri-toxicité, si ce n'est toti-toxicité.

Ainsi, l'objectif principal de ma thèse est d'élaborer une approche novatrice d'analyse et de traitement de données de toxicogénomique par l'implémentation de plusieurs méthodologies ayant traits aux (bio-)statistiques et à la (bio-)informatique dans la perspective de :

(i) discriminer des classes de substances chimiques par l'intégration de leurs signatures transcriptionnelles (c'est-à-dire l'ensemble des gènes dont l'expression est altérée positivement ou négativement à la suite d'une exposition à ce composé) ; Pour ce faire, la récupération exhaustive, autant que faire se peut, de l'ensemble des données de toxicogénomique présentes dans les banques publiques sera réalisée, afin de couvrir le plus de familles chimiques et de types de toxicité possibles.

(ii) prédire leurs effets toxicologiques par l'intégration des associations connues pour certains de ces composés avec des pathologies humaines et phénotypes délétères. Dans cet optique, les liens connus entre "gènes" et "pathologies" seront utilisés afin d'inférer, sans *a priori*, les toxicités potentielles des composés sur la base des gènes qu'ils affectent.

(iii) être ainsi capable de prioriser une liste de substances, sur la base de leurs signatures toxicogénomiques, en fonction d'un effet toxicologique recherché. L'ensemble de ces prédictions et priorisations permettra alors la mise en place de tests spécifiques et adaptés aux composés en fonction de leurs toxicités suspectées. Ceci devrait *in fine* faciliter l'évaluation et l'analyse de nouveaux risques sanitaires/environnementaux et ainsi contribuer indirectement à orienter les politiques publiques dans leur gestion. Plus particulièrement en ce qui concerne la thématique de recherche de mon laboratoire, mon projet contribuera ainsi à identifier et à valider de nouveaux perturbateurs endocriniens et/ou xénobiotiques reprotoxiques auxquels les humains sont exposés.

Comme précédemment effectué au sein de mon laboratoire avec un autre système de priorisation, GPSy (Gene Prioritization System) (Britto et al. 2012), les classes de composés obtenues et l'outil de priorisation associé, ChemPSy (Chemical Prioritization System), ont vocation à être mis à la libre disposition de la communauté scientifique. Ainsi, un deuxième objectif au cours de ma thèse est de développer une interface web, TOXsIgN (TOXicogenomics sIgNature database) constituant à la fois un espace de dépôt structuré pour les signatures toxicologiques et toxicogénomiques, une banque de données facilement interrogeable et permettant d'accéder aisément à l'ensemble des connaissances déposées, mais aussi un espace de travail mettant à disposition des outils pour l'analyse et la comparaison des données de toxicogénomique. En ce sens, ChemPSy constituera l'un des modules de TOXsIgN.

Enfin, un dernier objectif de mon doctorat est de déployer un navigateur de génome dédié à la communauté scientifique de la reproduction, RGV (ReproGenomics Viewer), afin d'y héberger et de faciliter l'accès à la quantité grandissante de données de séquençage (ChIP-seq, RNA-seq etc.) dans ce domaine. Ce "prototype" aura par ailleurs vocation à être transposé aux données de toxicogénomique afin de permettre la transition technologique également en cours dans ce domaine.

3. Matériels et méthodes

3.1. Ressources informatiques et environnement de travail

3.1.1. Langages de programmation

3.1.1.1. Python

Python est un langage de programmation multiplateforme orienté objet. Il s'utilise dans de nombreux contextes et peut s'adapter à tout type d'utilisateur grâce à une vaste bibliothèque d'outils allant de l'interfaçage de logiciel aux calculs numériques (simulation informatique par exemple). En raison de sa syntaxe facile (à lire et à comprendre), mais aussi pour sa rapidité de traitement des fichiers, le langage Python est utilisé principalement pour automatiser des tâches simples, mais fastidieuses et répétitives. Dans le cadre de ma thèse, Python a été utilisé pour le parcours de fichiers et le travail sur ces derniers (lecture, écriture, concaténation, expression régulière). En biologie, de nombreux outils ou *workflow* (enchaînement d'opérations) utilisent Python, par exemple l'outil de conversion de coordonnées entre les différentes versions de génome CrossMap (Zhao et al. 2014) que j'ai utilisé dans le cadre du projet RGV (cf. 4.3. The ReproGenomics Viewer).

3.1.1.2. Environnement R

R est un langage de programmation dédié aux statistiques et au traitement des données (R core Team 2013). Ce langage multiplateforme orienté objet est très utilisé en biologie notamment en génomique ou en épidémiologie pour la filtration statistique des données et leur représentation. Tout comme Python, R possède une vaste bibliothèque d'outils divers et variés parmi lesquels *Bioconductor* est la librairie la plus utilisée en bio-informatique puisqu'elle fournit de nombreux outils dédiés à l'analyse de données des sciences de la vie (Huber et al. 2015). Dans le cadre de mes travaux sur ChemPSy (cf. 4.1. ChemPSy), R a été utilisé pour effectuer le prétraitement et la filtration statistique des données de génomique, des analyses multivariées (ACP, ACM), de la classification non supervisée (Mclust, Hopach, Dynamic tree cut) ainsi que des analyses discriminantes (SVM, PLS).

Prétraitement et la filtration statistique des données de génomiques

Pour cette partie j'ai utilisé les packages RMA (*Robust Multi-Array Average*) (Irizarry et al. 2003) et Limma (*Linear Models for MicroArray*) (Ritchie et al. 2015). RMA, est une méthode corrigeant les valeurs d'intensité brute de puces à ADN après estimation et soustraction du bruit de fond. Les

intensités sont ensuite log2 transformées et normalisées par la méthode du quantile-quantile. L'étape finale de « *summarization* » permet ensuite d'estimer un niveau d'expression pour chaque transcrit représenté sur la puce. Limma (Ritchie et al. 2015) est un package d'analyse statistique de modèle linéaire permettant d'estimer la variance des données de puce à ADN ainsi que d'y appliquer une correction pour les tests multiples.

Analyses multivariées et analyses discriminantes

Sous R, l'ACP (cf. 3.4.1. Analyse en composantes principales) et l'ACM (Analyse des Correspondances Multiples) sont réalisées grâce à la librairie FactoMineR (Lê, Josse, and Husson 2008). C'est un outil dédié aux analyses exploratoires multidimensionnelles des données. Il permet entre autres d'effectuer des analyses statistiques multivariées. Ce module embarque également de nombreuses aides à l'interprétation (dans le cadre de l'ACP, une détermination automatique du nombre de composantes principales optimales), mais aussi des outils visuels pour la représentation graphique des résultats obtenus. Afin d'effectuer l'étape de prédiction, j'ai utilisé les méthodes de SVM et PLS (cf. 3.4.3. Méthodes de prédiction de classe). Sous R ces analyses ont été réalisées en utilisant les packages FactoMineR pour la PLS et le module libSVM (Bennett 2000) pour le SVM

Méthodes de classification

La première méthode de classification utilise la librairie **Mclust** (Chris Fraley, Adrian E. Raftery, T. Brendan Murphy 2012) pour le regroupement, la classification et l'estimation de densité basée sur la modélisation de mélange gaussien via l'algorithme d'espérance-maximisation. Il s'agit d'un algorithme itératif qui permet de trouver les paramètres du maximum de vraisemblance d'un modèle probabiliste lorsque ce dernier dépend de variables latentes non observables. La méthode de classification **Hopach** (*Hierarchical Ordered Partitioning And Collapsing Hybrid*) (van der Laan and Pollard 2003) a également été utilisée. Cette méthode force le partitionnement de chaque cluster formé en sous-clusters par l'utilisation de la méthode de la *Median Split Silhouette*. Cette méthode mesure l'hétérogénéité de chaque cluster. Si cette dernière est trop grande, Hopach divisera le cluster. Enfin la dernière méthode est celle de la librairie **Dynamic tree cut** (Langfelder, Zhang, and Horvath 2008). La méthode implémentée dans ce package permet de détecter les clusters au sein de dendrogrammes en fonction de sa forme.

3.1.1.3. Outils de version

Un outil de gestion de version (VCS) permet de suivre les modifications itératives réalisées sur un projet informatique. La possibilité de revenir à une version antérieure d'un projet ou d'un fichier facilite le développement à plusieurs. L'utilisation de VCS tel que GitHub ou SourceForge est recommandée dans le cadre du développement d'outils scientifiques (Prlić and Procter 2012). Ces logiciels donnent également accès à des listes de diffusion et de suivis des erreurs accessibles à toute la communauté scientifique utilisant le logiciel (Rother et al. 2012). Dans le cadre de ma thèse, l'outil de gestion de version GitHub a été utilisé (<https://github.com/>) pour l'ensemble de mes développements, mais également pour l'écriture de ce manuscrit de thèse.

3.1.2. Environnement web

La mise en place d'un outil web nécessite de mettre en relation différents composants afin de pouvoir gérer des données, les représenter de manière visuelle et d'exécuter les commandes effectuées par les utilisateurs sur le site web. Cette architecture est appelée architecture MVC (Modèle, Vue, Contrôleur) (Figure 23). Le projet TOXsIgN (cf. 4.2. TOXsIgN) a été développé selon ce modèle.

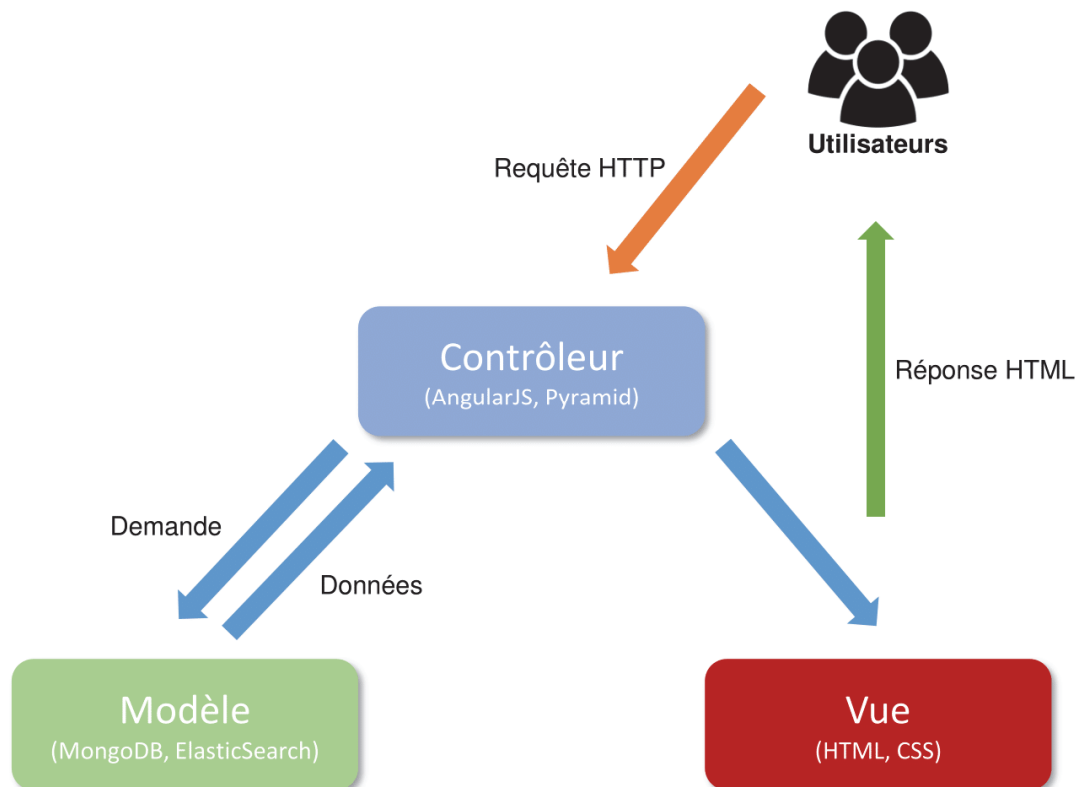


Figure 23. Schématisation du modèle MVC.

Le modèle contient les informations à afficher sur une page web, la vue est la représentation visuelle de la page et le contrôleur est le superviseur d'exécution des actions effectuées par l'utilisateur.

L'utilisation d'une telle architecture assure la structuration et la stabilité des sites web. Aujourd'hui, de nombreux sites ont adopté cette organisation (LinkedIn, Netflix, Youtube). Dans les projets web développés durant ma thèse, la structure MVC a été utilisée via l'utilisation des *frameworks* (outils facilitant le développement) web AngularJS et Pyramid. Les données ont été stockées et gérées grâce au système de gestion de base de données (SGBD) MongoDB et au serveur d'indexation et de recherche Elasticsearch. Enfin, l'interface web a été réalisée en utilisant les langages HTML5 et CSS3.

3.1.2.1. AngularJS

AngularJS (<https://angularjs.org/>) est un *framework* JavaScript libre et open source développé en 2009 par Google. Il est conçu autour de quatre grands concepts assurant la stabilité et la robustesse du site web. Le premier concept est celui précédemment évoqué de MVC répartissant de façon stricte les données, la présentation des données et les actions effectuées sur ces données. Le second concept est celui de Data Binding. Il s'agit ici du lien que fait AngularJS entre le code JavaScript (utilisé pour les fonctions et le traitement de certaines informations) et le code HTML (représentation de la page). Grâce à cela, il est possible pour chaque page HTML de créer un modèle propre et d'afficher des variables en temps réel sans traitement préalable en JavaScript. *Via* la gestion des dépendances, AngularJS permet à tout développeur de créer et de mettre à disposition ses propres modules AngularJS (outils). Enfin, le dernier concept est celui de la manipulation du DOM (*Domain Object Model*) qui correspond à la capture d'action faite sur une page web par exemple le clic sur un bouton, le scroll d'une page ou encore la gestion du clavier, de la page ou d'un formulaire. La capture de ces actions représente, le plus souvent, de nombreuses lignes de code difficilement maintenables et testables. AngularJS embarque des fonctions simples permettant de gérer rapidement ce type d'évènements.

3.1.2.2. Pyramid

Pyramid (<http://docs.pylonsproject.org/projects/pyramid/en/latest/#>) est un *framework* web python libre et open source. Il a été développé par *Pylons project*, une organisation de développeur ayant pour objectif de mettre en place des technologies web en utilisant le langage Python. La finalité de ce projet étant d'utiliser la simplicité d'utilisation et la facilité de lecture offertes par Python afin de mettre en place une architecture web robuste et aisément maintenable. Pyramid communique avec la banque de données et gère les requêtes effectuées depuis le site afin de fournir aux utilisateurs du site les informations demandées. Ce *framework* offre la possibilité d'utiliser toute la bibliothèque d'outils de Python ainsi que les outils ou scripts développés de manière indépendante.

3.1.2.3. MongoDB

MongoDB (<https://www.mongodb.com/fr>) est un système de gestion de base de données orientée document et évolutif de la mouvance NoSQL. Ceci signifie que les données stockées respectent un système clé-valeur où pour chaque information, une valeur est associée (« identifiant = adresse mail » par exemple). De grands groupes internationaux ont adopté le système NoSQL pour la gestion des leurs données, comme Google, Amazon ou encore Facebook.

Le côté évolutif de MongoDB, ou « *scalability* », se traduit par le fait que MongoDB est capable de s'adapter en fonction la capacité du dispositif informatique, mais également en fonction de la façon dont la base est sollicitée. L'utilisation d'objets au format binaire (BSON) sans schéma prédéterminé permet de manipuler les données de manière souple en les ajoutant et en les retirant de la base à tout moment « à la volée ».

Les données prennent la forme de documents enregistrés eux-mêmes dans des collections - une collection contenant un nombre quelconque de documents. Les collections sont comparables aux tables, et les documents aux enregistrements des bases de données relationnelles. Néanmoins, contrairement à ces dernières, les champs d'un enregistrement sont libres et peuvent être différents d'un enregistrement à un autre au sein d'une même collection. Le seul champ commun et obligatoire est le champ de clé principale ("_id"). Par ailleurs, MongoDB ne permet pas d'effectuer des requêtes très complexes, mais permet de programmer des requêtes spécifiques en JavaScript.

3.1.2.4. Elasticsearch

Elasticsearch (www.elastic.co) est un système de gestion de base de données NoSQL possédant un moteur de recherche. Ce dernier est utilisé pour indexer les données sous format texte. La rapidité et la fluidité des recherches au sein de la base de données sont possibles grâce à la présence d'une architecture parallèle et redondante. Elasticsearch est composé de plusieurs nœuds, chaque nœud étant une instance exécutée sur une machine séparée la plupart du temps et possédant une copie des informations indexée de la base de données. Cette architecture, déployant l'information en plusieurs copies sur des machines différentes, assure la tolérance aux pannes et offre la capacité de servir beaucoup de requêtes simultanément. L'organisation et la structure de données d'une entrée n'ont pas à suivre un schéma prédéfini et se présentent sous la forme d'un document JSON, encodage de texte de données structurées très utilisé pour les échanges entre processus distants. Ce format est également utilisé pour les résultats issus des requêtes effectuées *via* l'API REST (*Representational State Transfer*)

d'Elasticsearch permettant des opérations CRUD (Create, Read, Update, Delete) sur les données de la base.

Elasticsearch met à disposition de nombreux modes de recherche (*fuzzy like this, term, text, range...*). Lors des développements web réalisés durant ma thèse, le mode de recherche *QueryString* a été utilisé. Ce dernier utilise la syntaxe Lucene (Tableau III), bibliothèque open source écrite en Java sur laquelle Elasticsearch est basé, dans le but de construire des requêtes puissantes et précises. La requête « génomique^3 métabolomique –protéomique » par exemple, cherche tous les documents contenant les mots « génomique » et « métabolomique », mais sans le mot « protéomique », en affectant un meilleur score aux documents qui contiennent le mot « génomique », car on y a appliqué un « poids » de 3. Une fois la recherche effectuée, les résultats retournés sont évalués par pertinence.

Symbole	Description
AND	Opération ET, les deux termes doivent être présents
OR	Opération OU, les deux termes peuvent être présents
+	Le terme après doit obligatoirement être présent
-	Exclut les documents qui contiennent le terme
:	Recherche dans un champ spécifique
^	Augmente le poids d'un mot
*	Remplace un ou plusieurs caractères.
~	Coordonnées de début de la fonction par rapport au cadre de lecture
()	Regroupe plusieurs termes
?	Remplace un seul caractère

Tableau III.Symboles utilisés dans le moteur de recherche Elasticsearch

3.1.2.5. Hébergement des sites web

Les sites web développés au cours de ma thèse sont actuellement hébergés par la plateforme bio-informatique GenOuest sur des machines spécialement dédiées pour l'IRSET. Brièvement, ces machines se composent de 12 processeurs, 400 Go de mémoire vive et de 120 To de stockage. Chaque site web est encapsulé dans un conteneur Docker (<https://www.docker.com/>) qui permet d'empaqueter des applications et leurs dépendances dans des environnements Linux indépendants (conteneurs) du système sur lequel Docker est déployé. L'utilisation de ce système offre une plus grande sécurité et facilité de maintenance des sites web.

3.2. Outils et ressources bio-informatiques

3.2.1. Suites logicielles

Le système de conversion du *ReproGenomics Viewer* (cf. 4.3. RGV) repose sur l'enchaînement de plusieurs outils, qui successivement vont convertir le format de fichier initial, standardiser le fichier intermédiaire, convertir sur un autre génome les coordonnées du fichier standardisé et enfin assurer leur compatibilité avec le navigateur de génome.

3.2.1.1. Suite d'outils UCSC

La suite d'outils UCSC (Kuhn, Haussler, and Kent 2013) a été développée afin de préparer les fichiers à une visualisation dans le *genome browser* de l'UCSC. Elle permet un grand nombre de conversions de fichiers d'alignement et de fichiers de données quantitatives dans un format supporté par les navigateurs de génome. Cette suite a été utilisée afin de convertir des fichiers wig, bigWig ou bam en format bedgraph.

3.2.1.2. Suite de logiciels BedTools

BedTools (Quinlan 2014) est une suite de logiciels de traitement et de conversion de fichiers de données quantitatives (.bed). Dans RGV, nous avons utilisé la commande « merge » de BedTools afin de combiner toutes les entités biologiques chevauchantes dans les fichiers bedgraph et de moyenner leur valeur d'expression. Cette étape correspond à la standardisation des fichiers bedgraph.

3.2.1.3. CrossMap

CrossMap (Zhao et al. 2014) est un programme pour la conversion de coordonnées de génome. Contrairement à *liftover*, un outil similaire fourni par l'UCSC, CrossMap supporte de nombreux formats de fichier comme SAM / BAM, Wiggle / Bigwig, BED, GFF / FTE ou VCF. Dans RGV, il a été utilisé pour convertir les coordonnées entre les différents génomes ainsi qu'entre les différentes versions d'un même génome en utilisant les fichiers d'alignement par paires (*chaineFile*) fournis par UCSC.

3.2.2. Navigateurs de génome

Un navigateur génome ou *genome browser* est, comme son nom l'indique, un outil de visualisation de génome. Il permet via une interface dédiée de parcourir un génome en affichant diverses informations (ou pistes) en fonction de leur position sur le génome ce qui permet la comparaison visuelle

rapide de plusieurs jeux de données de génomique (par exemple transcriptomique et épigénomique) au sein de la même interface. Les utilisateurs peuvent cibler un chromosome entier, mais également une région spécifique par un système de grossissement afin d'accéder à des informations précises telles que la position d'un gène, de ses épissages alternatifs ou encore de sa séquence. La plupart des navigateurs permettent également aux utilisateurs de charger leurs propres données (dans des formats spécifiques tels que BED, WIG, BIGWIG, BAM) afin de faciliter le travail d'interprétation de l'utilisateur.

Il existe deux types de navigateurs de génome : les navigateurs généralistes et les navigateurs spécialisés. Par définition, un navigateur généraliste met à disposition diverses informations aussi bien qualitatives (localisation d'un gène codant), mais également des informations quantitatives (niveau d'expression d'un gène dans une condition expérimentale précise) pour plusieurs espèces, mais sans thématique biologique centrale. De nombreuses institutions proposent des navigateurs de génomes généralistes parmi lesquels ceux de Ensembl (<http://www.ensembl.org/index.html>) (Aken et al. 2017), du NCBI (<https://www.ncbi.nlm.nih.gov/genome>) (NCBI Resource Coordinators 2017) et de l'Université de Santa Cruz (UCSC) (<https://genome.ucsc.edu/>) (Kuhn, Haussler, and Kent 2013).

À la différence de ces navigateurs généralistes, d'autres que l'on qualifiera de « spécialisés » centralisent des informations d'une seule espèce (Choo, Heydari, et al. 2014; Heydari et al. 2014, 2013) ou d'une seule technologie (Hackenberg, Barturen, and Oliver 2011).

La démocratisation de ce type d'outils a permis à des consortiums de mettre à disposition de la communauté scientifique différents outils pour faciliter la mise en place de navigateurs personnalisés. Le *Integrated Genome Browser* (IGB) (Freese, Norris, and Loraine 2016) et *Integrative Genomics Viewer* (IGV) (Thorvaldsdottir, Robinson, and Mesirov 2013) offrent la possibilité d'installer un navigateur de génome sur son propre ordinateur. Ils permettent ainsi de visualiser, sans accès à internet, des informations génomiques pour plus de 100 espèces différentes. D'autres solutions permettent de mettre en place une interface web, comme *GBrowse* (Donlin 2009) et *Jbrowse* (Skinner et al. 2009), et offrent ainsi la possibilité de personnaliser les affichages, la visualisation, les génomes et les pistes à afficher. Dans le cadre du développement du *ReproGenomics Viewer* (cf. 4.3. RGV), l'utilisation de *Jbrowse* a été privilégiée.

3.2.3. Galaxy

Galaxy est une application web développée par l'université de Johns Hopkins (Giardine et al. 2005; Blankenberg et al. 2010). Elle offre aux biologistes la possibilité d'analyser des données issues de technologies à haut débit sans connaissance (bio-)informatique préalable. Cette abstraction permet aux

scientifiques de lever de nombreux verrous techniques (installation de logiciels) et ainsi de se concentrer sur leur question biologique et l'élaboration des stratégies d'analyse qui permettront d'y répondre. Cette application dispose d'une importante communauté de développeurs actifs mettant régulièrement le code source à jour et mettant à disposition de nombreux modules implémentant des logiciels libres utilisés en biologie dans la bibliothèque de Galaxy. Plus de 4800 outils sont ainsi disponibles permettant l'assemblage de génome, l'analyse de données métabolique ou encore la visualisation de données dans des navigateurs de génome.

L'autre avantage de Galaxy est de proposer un outil de création de *workflow*, c'est-à-dire de proposer une interface web dans laquelle il est possible de programmer un enchaînement d'outils à appliquer séquentiellement sur un jeu de donnée (Figure 24).

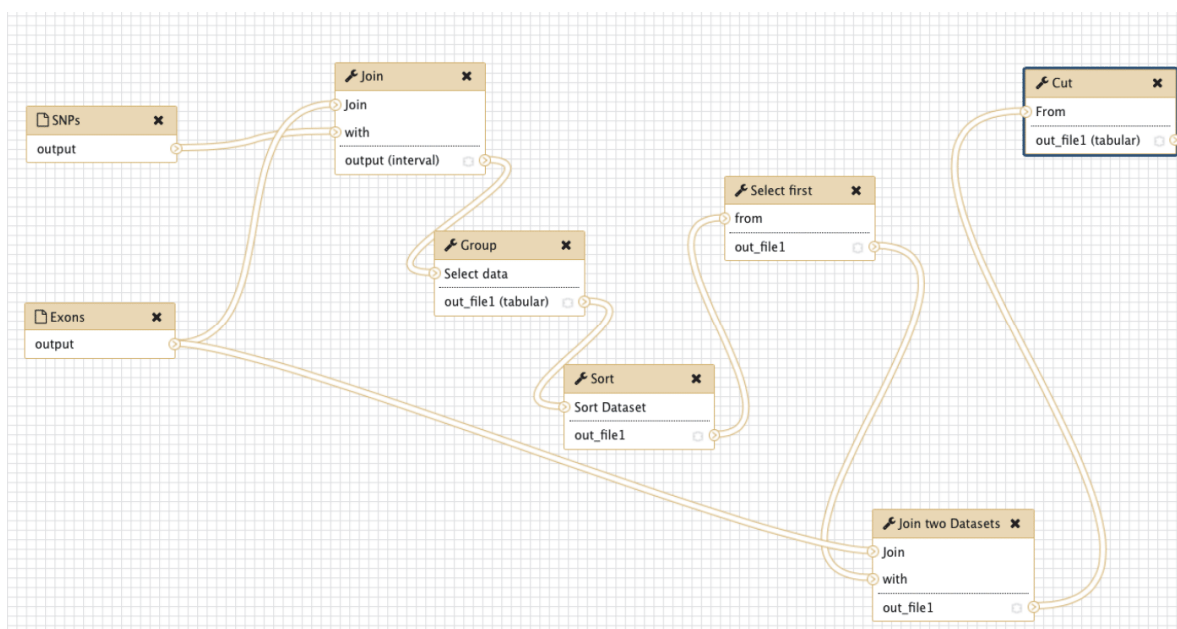


Figure 24. Exemple de workflow sous Galaxy.

Chaque outil est représenté par une boîte et chaque boîte est associée par une flèche représentant l'ordre d'utilisation des programmes pour l'analyse des données d'entrées

Dans un objectif clairement affiché de « réaliser une science génomique reproductible et transparente » (Orvis et al. 2010), Galaxy facilite également l'échange de workflows et de procédures d'analyses à appliquer sur les données brutes décrites dans les publications, afin de donner un accès simplifié aussi bien aux outils utilisés qu'aux paramètres définis pour obtenir les résultats décrits dans un papier.

Il existe plusieurs serveurs Galaxy accessibles par tous avec ou sans inscription, la plus utilisée étant probablement celle de la Galaxy Team (<https://usegalaxy.org/>). De nombreuses plateformes bio-informatiques ont également mis en place leurs propres instances (Le Bras et al. 2013) ou ont mis à disposition des instances centralisées sur un domaine en particulier (Sundell et al. 2015) comme l'espace de travail du *ReproGenomics Viewer* par exemple (cf. 4.3. RGV).

3.2.4. Vocabulaires contrôlés et ontologies

La création des ontologies est le résultat textuel d'un effort commun fait dans le domaine biomédical pour uniformiser le vocabulaire utilisé. Ce vocabulaire contrôlé, ou ontologie est un ensemble structuré de termes et/ou de concepts ordonnés de manière hiérarchique (parents-enfants) représentant les éléments d'un domaine de connaissances par exemple la taxonomie (NCBI Resource Coordinators 2017), l'anatomie humaine (Rosse and Mejino 2003) ou les composés chimiques (Hastings et al. 2013). L'ontologie structure le vocabulaire utilisé dans l'un de ces domaines et ainsi uniformise le champ lexical utilisé. Face au besoin croissant de structuration des informations en générale, des bases de données centralisant ces ontologies, telles que OBO Foundry (Smith et al. 2007) ou Ontobee (Zuoshuang et al. 2011).

Afin d'ordonner et de structurer les données de toxicologie déposées dans TOXsIgN (cf. 4.2. TOXsIgN) et ChemPSy (cf. 3.5. Méthodes appliquées dans le cadre du projet ChemPSy), l'utilisation d'un vocabulaire contrôlé a été mise en place. Pour chaque objet décrit, une ontologie spécifique est associée. Les espèces sont décrites via l'ontologie NCBI Taxonomy (NCBI Resource Coordinators 2015). Cette ontologie référence tous les organismes présents dans les bases de données publiques. L'utilisation de l'ontologie ChEBI (Degtyarenko et al. 2009) permet de représenter les molécules utilisées lors des expérimentations. ChEBI se présente comme la ressource publique des entités moléculaires, avec un accent plus particulier sur les composés chimiques de 'petite taille'. La description du design expérimental est effectuée par l'ontologie OBI (Ontology for Biomedical Investigations) (Brinkman et al. 2010), qui décrit les expérimentations biologiques et cliniques. Les cellules, lignées cellulaires et tissus utilisés sont décrits par les ontologies Cell Ontology (Bard, Rhee, and Ashburner 2005), Cell Line Ontology (Sarntivijai et al. 2014), Foundational Model of Anatomy (Rosse and Mejino 2003) et BRENDA (Gremse et al. 2011). Brièvement, Cell Ontology est un vocabulaire contrôlé pour la description des types cellulaires, Cell Line Ontology permet de normaliser et d'intégrer des informations relatives aux lignées cellulaires, Foundational Model of Anatomy est une ontologie descriptive de l'anatomie humaine, tandis que BRENDA est un vocabulaire contrôlé et structuré pour décrire les tissus ainsi que les lignées cellulaires, types cellulaires et cultures cellulaires. Les processus biologiques sont

décrits par le sous-ensemble « Biological process » de la Gene Ontology (Ashburner et al. 2000). Gene Ontology est un projet bio-informatique destiné à structurer la description des gènes et des produits géniques de manière commune à toutes les espèces.

3.2.5. Bases de données

3.2.5.1. Banques de données génomiques

Gene Expression Omnibus (GEO)

GEO (Barrett et al. 2013) (<https://www.ncbi.nlm.nih.gov/geo/>) est un espace web offrant aux utilisateurs la possibilité de déposer des données brutes « omiques », telles que des données de transcriptomique. Il regroupe un grand nombre d'échantillons (2087591 en juin 2017) représentant la base de données de choix pour constituer le jeu de données de ChemPSy ChemPSy (cf. 3.5 Méthodes appliquées dans le cadre du projet ChemPSy).

Entrez gene

Entrez gene (<https://www.ncbi.nlm.nih.gov/gene>) est une base de données maintenue par le NCBI centrée autour des gènes (Maglott et al. 2011). Chaque gène est associé à un identifiant unique ainsi qu'à de nombreuses informations par exemple le nom, les synonymes, l'espèce ou encore les produits de gène (transcrits, protéines). Les Entrez gene IDs étant liés à un grand nombre de banques de données, ils permettent de s'affranchir de la barrière technologique – la majorité des entités biologiques (gènes, transcrits, protéines, sondes nucléotidiques, ...) pouvant être converties en Entrez gene IDs. Les informations de cette banque sont utilisées dans l'outil de comparaison de signatures de TOXsIgN (cf. 4.2. TOXsIgN).

HomoloGene

HomoloGene (<https://www.ncbi.nlm.nih.gov/homologene/>) (NCBI Resource Coordinators 2016) est une banque d'homologie développée par le NCBI. Elle compare des séquences protéiques les unes par rapport aux autres et les regroupe au moyen d'arbres construits sur la similarité de ces séquences. La dernière version d'HomoloGene indexe 21 espèces. Comme pour la banque Entrez gene, HomoloGene est utilisée dans l'outil de comparaison de signatures toxicogénomiques TOXsIgN (cf. 4.2. TOXsIgN).

UCSC

La base de données de l'UCSC (*University of California Santa Cruz*) (<https://genome.ucsc.edu/index.html>) met à disposition des nombreuses ressources génomiques et outils bio-informatiques directement téléchargeables depuis un serveur FTP. Parmi ces ressources, les séquences et annotations du génome de plus d'une centaine d'espèces et de génomes sont disponibles. Dans le cadre du projet RGV (cf. 4.3. RGV), les séquences et annotations de génomes d'intérêt ont été téléchargées depuis cette ressource.

3.2.5.2. Bases de données toxicologiques

La Comparative Toxicogenomics Database (CTD)

La CTD (<http://ctdbase.org/>) ou Comparative Toxicogenomics Database (A. P. Davis et al. 2015) est une banque de données spécialisée dans le domaine de la toxicogénomique maintenue par le département des sciences biologiques de l'université de Caroline du Nord. Elle met en relation les composés chimiques, les gènes et les pathologies sur la base d'associations mises en évidence dans la littérature. À ces associations, des informations sur les voies métaboliques perturbées, les phénotypes induits ainsi que sur l'annotation des gènes sont également fournis. Cette banque de données est utilisée dans le cadre du projet ChemPSy (cf. 3.5 Méthodes appliquées dans le cadre du projet ChemPSy) et TOXsIgN (cf. 4.2. TOXsIgN) pour les fichiers d'interactions composés-gènes, composés-pathologies et l'ontologie sur les pathologies, tous téléchargeables depuis le site de la CTD.

DrugBank

DrugBank (Law et al. 2014) (<https://www.drugbank.ca/>) est une base de données publique centrée sur la bio-informatique et la chémo-informatique. Cette base est hébergée par l'université de l'Alberta, Canada. Aujourd'hui, DrugBank contient des informations sur 8261 médicaments, dont 2021 approuvés par la FDA et environ 6000 médicaments expérimentaux. Toutes ces informations sont associées à près de 4 300 séquences protéiques. Cette base de données est utilisée dans ChemPSy (cf. 3.5 Méthodes appliquées dans le cadre du projet ChemPSy) afin d'obtenir plus d'informations sur les composés chimiques du jeu de données.

3.2.6. Description des différents types de fichiers

Durant mon travail de thèse, un grand nombre de fichiers différents a été utilisé. Par la suite j'utiliserai uniquement l'extension de ses fichiers pour les citer, il convient de décrire le format des informations stockées dans ces fichiers, mais également où et comment ils sont utilisés.

3.2.6.1. Fasta et Fastq

Le fichier fasta est un format de fichier utilisé en biologie pour stocker une ou plusieurs séquences nucléotidiques ou protéiques. Il est au minimum composé de deux lignes et se présente sous la forme suivante :

La première ligne contient un identifiant de séquence situé juste après le symbole « > », d'autres informations peuvent également être indiquées sur la suite de la ligne. Toutes les autres lignes correspondent aux résidus (nucléotidiques ou protéiques). Ce format est donc couramment utilisé pour décrire la structure primaire d'une séquence et constitue souvent le fichier d'entrée de nombreux outils permettant, par exemple, d'aligner des séquences ou de réaliser des prédictions de motifs.

Le format fastq dérive du fichier fasta et permet de représenter une ou plusieurs séquences avec leurs scores de qualité par base. Il utilise en principe 4 lignes par séquence. La ligne 1 commence par un caractère "@" suivi de l'identifiant de la séquence et éventuellement d'une description (de la même façon qu'un fichier au format FASTA, le "@" remplaçant ici le ">"). La ligne 2 contient la séquence nucléique brute. La ligne 3 commence par un caractère "+", parfois suivi par la répétition de l'identifiant de la séquence et de sa description si celle-ci est présente. La ligne 4 contient les scores de qualité associés à chacune des bases de la séquence de la ligne 2 et doit avoir exactement le même nombre de symboles que la ligne 2.

3.2.6.2. SAM et BAM

Le format SAM (*Sequence Alignment/Map*) est un format d'alignement générique utilisé pour stocker les alignements de lectures (reads, séquences) sur des séquences de référence. Le format SAM est assez flexible pour stocker toutes les informations d'alignements produites par divers programmes d'alignement. Mais il est également compact et peut être indexé en fonction des positions génomiques afin d'augmenter la vitesse de récupération d'informations sur ce type de fichier. Le format BAM est un format de fichier binaire compressé stockant des données de séquences issues du fichier SAM.

3.2.6.3. Bedgraph et Wiggle

Les formats bedgraph (BED) et wiggle (WIG) sont très utilisés dans la plupart des navigateurs de génome, car ils permettent de stocker simplement des régions génomiques auxquelles sont associées des informations qualitatives ou quantitatives telles que le pourcentage en GC, des scores de probabilité ou encore des abondances d'expression. La manipulation de ces fichiers rend le travail sur les séquences génomiques (comparaison par exemple) plus efficace en utilisant des coordonnées afin d'extraire des

séquences d'intérêt de jeux de séquençage ou de comparer et manipuler directement deux jeux de coordonnées. Divers programmes permettent de manipuler ces fichiers tels que BedTools (Quinlan 2014) et BEDOPS (Neph et al. 2012).

Un fichier BED se présente sous forme tabulée et est composé de trois, quatre, six ou 12 colonnes en fonction de la précision d'annotation (Tableau IV).

Colonne	Nom	Description
1	chrom	Nom des chromosomes (ex. : chr3, chrY)
2	chromStart	Coordonnée de départ sur le chromosome
3	chromEnd	Coordonnée de fin sur le chromosome
4	name	Nom de la ligne
5	score	Valeur numérique
6	strand	Sens (positif ou négatif)
7	thickStart	Coordonnée de départ à partir de laquelle l'annotation est affichée de façon plus épaisse sur une représentation graphique (ex. : le codon start d'un gène)
8	thickEnd	Coordonnées de fin à partir de laquelle l'annotation n'est plus affichée de façon plus épaisse sur une représentation graphique (ex. : le codon stop d'un gène)
9	itemRgb	Valeur déterminant la couleur d'affichage de l'annotation contenue dans le fichier BED
10	blockCount	Nombre de blocs (ex.: exons) sur la ligne du fichier BED
11	blockSizes	Liste de valeurs séparées par des virgules correspondant à la taille des blocs (le nombre de valeurs doit correspondre à celui du blockCount)
12	blockStarts	Liste de valeurs séparées par des virgules correspondant aux coordonnées de départ des blocs, calculées relativement à celles présentes dans la colonne chromStart (le nombre de valeurs doit correspondre à celui du blockCount)

Tableau IV.Colonnes composant un fichier BED

Enfin, le format WIG peut être trouvé sous forme compressée binaire appelée bigWig.

3.2.6.4. GFF et GTF

Le fichier GFF pour *General Feature Format* est un format standardisé et tabulé pour stocker des informations génomiques sous la forme de neuf colonnes (Tableau V). Il est fréquemment utilisé pour l'échange de données et la représentation de données génomiques dans des navigateurs de génome, par exemple. Il existe plusieurs versions de GFF dont la version 3, dernière version disponible pour ce format, qui est la version la plus utilisée aujourd'hui.

Le format GTF correspond en réalité la version 2.5 du format GFF. Celui-ci ajoute une fonctionnalité de représentation de plusieurs niveaux hiérarchiques selon différentes caractéristiques ou

fonctionnalités limitées par un vocabulaire contrôlé. Par exemple, il est possible de définir les coordonnées génomiques des introns et des exons appartenant à un transcrit, lui-même appartenant à un gène.

Colonne	Nom	Description
1	seqname	Nom du chromosomes (ex. : chr3, chrY)
2	source	Nom du programme ou de la source dont est issu le fichier
3	feature	Fonction / Caractéristique (ex : gène, répétition, variation...)
4	start	Coordonnée de départ de la fonction sur le chromosome
5	end	Coordonnée de fin de la fonction sur le chromosome
6	score	Valeur numérique
7	strand	Sens (positif ou négatif)
8	frame	Coordonnées de début de la fonction par rapport au cadre de lecture
9	attribute	Liste d'attributs de fonctionnalités séparés par un point-virgule. Se présente sous forme de paires de la manière suivante : « tag = valeur » (ex : ID = gene00001)

Tableau V.Colonnes composant un fichier GFF

3.2.6.5. CEL et CDF

Le format propriétaire CEL est dédié à la description des puces à ADN Affymetrix, ou GeneChip. Brièvement, ces fichiers sont créés par le logiciel d'analyse d'image micro-image d'ADN d'Affymetrix et contiennent une valeur d'intensité pour les millions de sondes nucléotidiques présentes sur la puce. Pour lire et extraire une valeur d'intensité normalisée pour chaque transcrit, il est nécessaire d'associer ces fichiers CEL à un fichier CDF qui décrit l'organisation de la puce utilisée et ainsi d'associer les sondes à leurs gènes/transcrits respectifs.

3.2.6.6. OBO et OWL

Le format OBO (*Open Biomedical Ontologies*) stocke, pour une ontologie donnée, les informations relatives à l'ordonnancement du vocabulaire et également la description et les synonymes pour chaque terme présent dans le fichier. Le format OWL (*Ontology Web Language*) est un langage de représentation des ontologies web. Le langage OWL est basé sur les recherches effectuées dans le domaine de la logique de description. Il peut être vu en tant qu'un standard informatique qui met en œuvre certaines logiques de description et permet à certains outils capables de comprendre OWL de travailler avec ces données, de vérifier que les données sont cohérentes, de déduire des connaissances nouvelles ou d'extraire certaines informations de bases de données.

3.2.6.7. JSON

Le format JSON ou *JavaScript Object Notation* est un format d'échange de données issu du langage JavaScript. Un document JSON a pour fonction de représenter de l'information selon un ensemble de paires clé/valeur sans aucune restriction sur le nombre de celles-ci. Ce format de fichier permet de représenter à la fois des objets, tableaux ou des valeurs génériques. Par exemple, ce format est utilisé pour l'échange d'informations entre la banque de données et l'affichage du site internet TOXsIgN (cf. 4.2. TOXsIgN).

3.3. Estimateurs mathématiques

3.3.1. Estimateurs de distances

Dans le cadre du projet ChemPSy, afin d'estimer la proximité entre les signatures toxicogénomiques, plusieurs distances ont été utilisées.

3.3.1.1. La distance euclidienne (DE)

Cette distance est probablement la plus couramment utilisée. Il s'agit d'une simple distance géométrique dans un espace multidimensionnel entre un point A (x_A, y_A) et un point B (x_B, y_B) calculé par l'équation suivante :

$$dAB = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2}$$

Le log de la distance euclidienne permet dans notre cas de centrer et de réduire les valeurs.

3.3.1.2. La corrélation de Pearson

Cette distance, également appelée coefficient de corrélation linéaire, mesure la relation linéaire (ou proportionnalité) entre deux séries de valeurs. Une droite de régression ou droite des moindres carrés est calculée pour minimiser la somme des carrés des distances de tous les points à la droite. La corrélation de Pearson suppose que les deux variables sont mesurées sur au moins une échelle d'intervalle. Le coefficient de corrélation de Pearson entre deux variables X et Y est calculé suivant l'équation :

$$r = \frac{COV(x, y)}{\sigma_x \sigma_y}$$

où $COV(x, y)$ est la covariance des variables x et y et σ_x et σ_y sont les écarts types de x et de y. Ce coefficient varie entre -1 et 1. Un coefficient de 1 indique une corrélation positive parfaite entre les deux variables. À l'inverse, un coefficient de -1 indique une corrélation négative parfaite, on parle d'anti-corrélation. Dans les deux cas, les points tombent parfaitement sur la droite. Un coefficient de 0 indique qu'il n'y a aucune relation entre les deux variables - la variation de l'une n'est pas associée à la variation de l'autre.

3.3.1.3. Le Hubert's score

C'est une méthode statistique d'estimation du degré d'association entre deux entités. Il s'agit ici d'une méthode alternative à l'odds ratio (présenté plus bas) pour l'évaluation de la co-citation entre deux

individus dans la littérature (la plupart du temps les gènes). L'avantage de cette méthode est qu'elle prend en compte le nombre d'associations entre les deux individus, mais également la force de cette évidence. Le Hubert's score prend en compte les paramètres du Tableau VI, adaptés à nos matrices discrétisées :

	Gène A up (1)	Gène A down(-1)	Pas de changement (0)
Gène B up (1)	n_{11}	n_{01}	n_{01}
Gène B down (-1)	n_{10}	n_{11}	n_{01}
Pas de changement (0)	n_{10}	n_{10}	n_{00}

Tableau VI.Paramètres adaptés aux matrices discrétisées de ChemPSy

Adapté aux matrices discrétisées le calcul du Hubert's score est :

$$\frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11} + n_{10})(n_{01} + n_{00})(n_{11} + n_{01})(n_{10} + n_{00})}}$$

3.3.1.4. L'indice de Kappa

L'indice de Kappa ou Kappa de Cohen est un test non paramétrique variant entre 0 et 1. Il est utilisé pour évaluer le degré de concordance entre deux individus. Ce qui le différencie du simple calcul d'une proportion (rapport entre le nombre d'éléments identiquement classés et le nombre total d'éléments à classer), est le fait qu'il introduit une correction pour prendre en compte le fait qu'une certaine proportion d'accord peut être imputée au hasard. L'indice traduit un niveau d'accord (de concordance) d'autant plus élevé que sa valeur est proche de 1. Cet indice se calcule par la formule suivante :

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)}$$

3.3.1.5. L'index de Jaccard

L'index de Jaccard est un indice utilisé en statistique pour comparer la similarité entre des échantillons. En partant de la matrice utilisée pour le calcul du Hubert's score, cet indice est défini par le nombre de gènes sur- et sous-exprimés dans les deux conditions à comparer divisées par la taille de l'union des deux conditions :

$$n_{11} + n_{10} + n_{01} + n_{00} = n$$

$$J = \frac{n_{11}}{n_{01} + n_{10} + n_{11}} = \frac{n_{11}}{n - n_{00}}$$

3.3.1.6. Le rapport des cotes

L'**odds ratio** ou rapport des cotes correspond au rapport de la probabilité d'apparition de l'événement dans un groupe traité divisé par la probabilité d'apparition de l'événement dans le groupe contrôle. Dans le cas de nos matrices discrétisées, le ratio est calculé comme suit :

$$OR = \frac{n_{11}n_{00}}{n_{10}n_{01}}$$

3.3.2. Évaluation d'un modèle statistique de classification

La validation croisée ou « cross-validation » constitue une méthode standard de mesure des performances d'un modèle. Il existe 3 grandes méthodes de cross validation : Holdout, LOOCV (leave-one-out cross-validation) et k-fold. La méthode utilisée pour la validation des modèles construits dans le cadre de ChemPSy est la méthode Holdout (également la méthode la plus utilisée). L'objectif de cette méthode est de séparer un ensemble de données en deux sous-ensembles. Le premier est un sous-ensemble de données destiné à l'apprentissage du modèle. Le second sert à tester ce modèle afin de l'évaluer. Le sous-ensemble de données d'apprentissage est très généralement plus grand que celui de test. On constate une proportion de 80% pour l'apprentissage et de 20% pour les tests. Un modèle statistique de classification sert à classer (étiqueter) de nouveaux individus non classés. Pour créer ce modèle, des individus appartenant à des classes déjà définies sont utilisés. Afin de tester la qualité de ce modèle, certains des individus sont mis de côté et étiquetés pour des « tests ». Ainsi, les individus « tests » sont classés sans a priori en les passant dans le modèle. Enfin, ces nouvelles étiquettes sont comparées avec les vraies étiquettes pour connaître le taux de bonne classification et donc la qualité du modèle. En croisant, les étiquettes réelles des individus avec leur nouvelle étiquette prédites par le modèle, on va pouvoir mettre en place une matrice de confusion servant à mesurer la qualité d'un système de classification (Tableau VII). Un des intérêts de la matrice de confusion est qu'elle montre rapidement si le système parvient à classer correctement.

		Classe estimée		
		Groupe 1	Groupe 2	Groupe 3
Classe réelle	Groupe 1	95	6	4
	Groupe 2	3	97	5
	Groupe 3	8	7	90

Tableau VII. Matrice de confusion

La matrice de confusion est le résultat de la reclassification d'individus dans un ensemble de classes. Elle compare la position d'individus par rapport à sa classe d'origine démontrant si le système parvient à classer correctement.

Pour analyser plus finement la qualité des classes produites par le modèle, la table de confusion de chaque classe est utilisée. Une table de confusion est une matrice de confusion organisée différemment où les classes non analysées sont regroupées en une seule et unique classe. Ainsi pour chaque classe, les individus sont répartis de façon binaire : dans la classe, pas dans la classe (Tableau VIII).

A

	Groupe 1	Pas groupe 1
Test positif	95	10 (6+4)
Test Négatif	11 (8+3)	199 (90+97+5+7)

B

	Classe	Non classe
Test positif	TP	FP
Test Négatif	FN	TN

Tableau VIII. Tableau de confusion

Le tableau de confusion est une matrice de confusion organisée autour d'une classe particulière. Elle classe les individus en fonction de leur appartenance ou non au groupe étudié. Par exemple, la table A, se focalise sur le groupe 1 et rapporte le classement des individus appartenant à ce groupe (test positif) ou non (test négatif) dans le groupe ou pas. La table B, correspond aux noms donnés aux valeurs de la table de confusion : vrais positifs (TP, *True Positive*) : nombre d'individus de la classe bien prédits dans celle-ci. Faux positifs (FP, *False Positive*) : nombre d'individus prédits dans la classe alors que ceux-ci n'y appartiennent pas. Faux négatifs (FN, *False Negative*) : individus prédits comme appartenant au groupe alors qu'ils ne devraient pas. Vrais négatifs (TN, *True Negative*) : individus n'appartenant pas au groupe et n'étant pas classés dans le groupe.

La généralisation de ce tableau permet de calculer pour chaque classe le nombre de vrai positif, faux positif, vrai négatif et faux négatif. Les vrais positifs (TP, *True Positive*) correspondent au nombre d'individus de la classe bien prédits dans celle-ci. Les faux positifs (FP, *False Positive*) correspondent

au nombre d'individus prédits dans la classe alors qu'ils n'y appartiennent pas. Les faux négatifs (FN, *False Negative*) sont les individus prédits comme appartenant au groupe alors qu'ils ne devraient pas. Les vrais négatifs (TN, *True Negative*) sont les individus n'appartenant pas au groupe et n'étant pas classé dans le groupe.

À partir de ces informations, il est possible de calculer différentes mesures pour évaluer le modèle : le taux d'erreur, la sensibilité et la spécificité. Le taux d'erreur correspond à la qualité générale du modèle. Ce taux est obtenu par la division des bonnes prédictions par le nombre total de prédictions, le tout soustrait de 1 :

$$\text{Taux d'erreur} = 1 - \left(\frac{\text{Nombre de bonne prédiction}}{\text{Nombre total de prédiction}} \right)$$

Dans notre cas, pour chaque individu reclassé, une liste de groupe ordonnée par ordre de classification est obtenue. On considérera comme bien prédit un individu dont le groupe se trouve dans le top n de la liste ordonnée, où n est un nombre entier compris entre 1 et le nombre de groupe du modèle.

La sensibilité (ou sélectivité) d'un test mesure sa capacité à donner un résultat positif lorsqu'une hypothèse est vérifiée. Elle est calculée par la formule suivante :

$$\text{Sensibilité} = \frac{TP}{(TP + FN)}$$

La sensibilité s'oppose à la spécificité, qui mesure la capacité d'un test à donner un résultat négatif lorsque l'hypothèse n'est pas vérifiée. Elle est définie par :

$$\text{Spécificité} = \frac{TN}{(TN + FP)}$$

3.4. Méthodes statistiques

Le projet ChemPSy (cf. 3.5 Méthodes appliquées dans le cadre du projet ChemPSy) repose sur plusieurs méthodes statistiques dont le fonctionnement théorique est présenté dans cette section.

3.4.1. Analyse en composantes principales

L'analyse en composante principale ou ACP consiste à transformer un jeu de variables quantitatives corrélées entre elles en un nouveau jeu de variables, moins nombreuses, mais indépendantes appelées composantes principales. Ces nouvelles variables diminuent la dimension du système tout en minimisant la perte d'information. Dans l'ACP, l'ensemble des données est considéré comme un espace à n dimensions où n est le nombre de variables du jeu de données. Chaque échantillon du jeu de données est représenté par un point dans cet espace à n dimensions $[x_1, x_2, \dots, x_n]$. La première étape de cette méthode statistique va construire une première composante principale PC1 de coordonnées t_1 , correspondant à la combinaison linéaire des n variables passant au plus près de tous les points du système :

$$t_1 = a_{1,1}x_1 + a_{2,1}x_2 + \dots + a_{n,1}x_n$$

où $a_{i,1}$ est une constante choisie pour maximiser la variance sur t_1 .

Sur le même principe, une deuxième composante principale PC2 de coordonnée t_2 est construite de manière orthogonale à la première. PC1 et PC2 sont ainsi indépendantes.

$$t_2 = a_{1,2}x_1 + a_{2,2}x_2 + \dots + a_{n,2}x_n$$

où $a_{i,2}$ est choisie de manière à ce que la variance sur t_2 soit maximale tout en imposant l'orthogonalité des vecteurs $[a_1]$ et $[a_2]$. La procédure est exactement identique pour tout ajout de nouvelles composantes. Plus généralement on notera chaque composante de la manière suivante : $t_h = Xa_h$

La première composante principale représente la plus grande part de la variance globale du système étudié. Les composantes successives expliquent uniquement la variance résiduelle du système. La détermination du nombre de composantes principales suffisantes pour expliquer la variance du système requiert une démarche de validation interne. Une composante PC_x est considérée comme utile si celle-ci contribue de manière significative à améliorer la robustesse de l'analyse.

La plupart du temps l'ACP est représentée en deux dimensions. Elle s'effectue en suivant les composantes PC1 et PC2 qui sont les deux composantes caractérisant la majeure partie de la variance dans le système. Le système peut alors être visualisé suivant deux matrices différentes (Figure 25). La première est une matrice de coordonnées, permettant d'analyser la dispersion des individus dans l'espace défini par les deux composantes majeures. Dans ce cas, deux individus proches graphiquement porteront une information très similaire. Cette proximité est calculée à l'aide de la distance euclidienne. La seconde matrice, représente une matrice des poids des différentes variables dans les composantes principales. Dans cette représentation, les variables contribuant le moins vont s'agglomérer à proximité du point d'origine, celles contribuant le plus vers les valeurs (positives ou négatives) importantes. Le calcul de la proximité entre ces variables se fait par le calcul de la corrélation entre les composantes principales et les variables. Cette corrélation est ensuite représentée dans un cercle des corrélations, correspondant à la projection du nuage des variables sur le plan des composantes principales. Les variables bien représentées sont celles qui sont proches du cercle, celles qui sont proches de l'origine sont mal représentées.

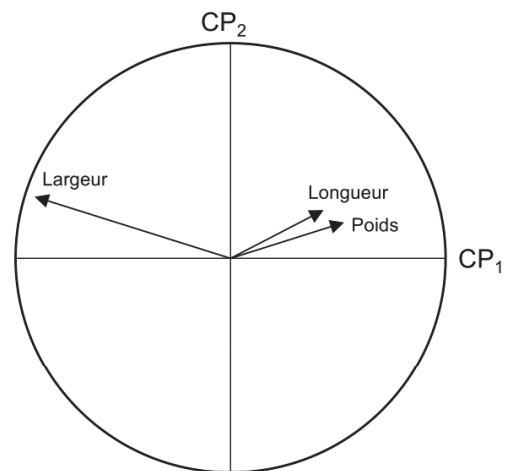
L'ACP est donc une méthode d'analyse non supervisée utile dans l'identification de variables fondamentales ou de groupes de variables corrélées diluées dans des jeux de données volumineux. Son utilisation principale, outre la réduction du jeu de données, est la mise en avant des caractéristiques du système étudié offrant la possibilité de trier graphiquement les individus par classe ou de visualiser les variables corrélées.

Donnée brutes

Poids	Longueur	Largeur
25	0.55	0.50
26	0.70	0.20
22	0.32	0.80
15	1	1.20

Coordonnées des variables

	CP ₁	CP ₂	CP ₃
Poids	0.99	0.15	0.0
Longueur	0.82	0.57	0.0
Largeur	-0.94	-0.34	0.0



Coordonnées des individus

	CP ₁	CP ₂	CP ₃
i ₁	-0.7	-0.75	0.0
i ₂	0.91	0.70	0.0
i ₃	-0.43	0.16	0.0
i ₄	-0.61	0.29	0.01

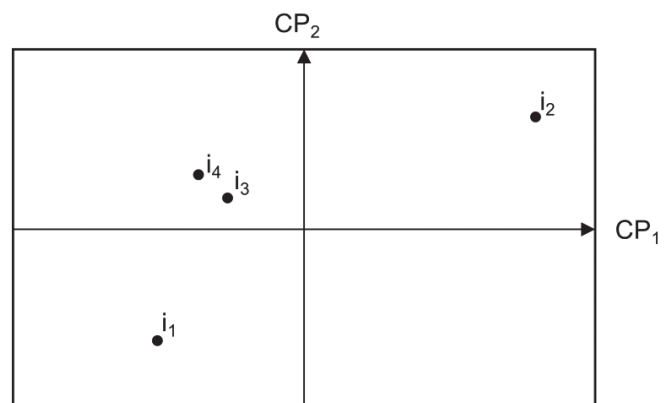


Figure 25. Représentation en deux dimensions de l'ACP : matrice de coordonnées et de poids.

Représentation des variables et des individus après ACP à partir d'une matrice de données brutes mettant en relation le poids, la longueur et la largeur. Les variables poids et longueur sont corrélées et expliquent moins la composante principale 1 (CP₁) que la largeur. Enfin la faible distance entre les individus i₃ et i₄ indique que ces derniers portent des informations similaires.

Durant mes travaux de thèse, j'ai également utilisé, dans une moindre mesure, l'analyse en correspondance multiple ou ACM. Brièvement, l'ACM est une méthode de description statistique multidimensionnelle de données qualitatives. Comme pour l'ACP, l'ACM permet soit de représenter de manière graphique le contenu d'un tableau de données, par similitudes entre les individus ou par modalités des variables qualitatives, soit de recoder en données numériques le jeu de données afin d'appliquer sur ce dernier d'autres méthodes statistiques. À la différence de l'ACP, l'ACM n'utilise pas les données brutes, mais prétraite ces informations. Pour cela cette méthode utilise en entrée un tableau disjonctif complet (TDC) avec en ligne les individus et en colonnes les modalités ou variables. Cette méthode peut également s'effectuer sur un tableau de Burt. Un TDC est une représentation de données qualitatives où une variable représentée par K modalités est remplacée par K variables binaires, chacune correspondant à une modalité (Figure 26). Une table de Burt est une matrice carrée symétrique contenant tous les tableaux de contingence des variables prises deux à deux. Enfin, l'ACM utilise la distance du χ^2 au lieu de la distance euclidienne pour calculer l'écart entre deux individus.

Individu	Sexe	Yeux
Père	Masculin	Marron
Mère	Féminin	Bleu
Enfant	Masculin	Vert



A

Individu	Sexe F	Sexe M	Yeux B	Yeux M	Yeux V
Père	0	1	0	1	0
Mère	1	0	1	0	0
Enfant	0	1	0	0	1

B

	Sexe F	Sexe M	Yeux B	Yeux M	Yeux V
Sexe F	1	0	1	0	0
Sexe M	0	2	0	1	1
Yeux B	1	0	1	0	0
Yeux M	0	1	0	1	0
Yeux V	0	1	0	0	1

Figure 26. Tableau disjonctif complet et tableau de Burt

A- Tableau disjonctif complet (TDC). C'est une représentation de données qualitatives où une variable représentée par K modalités est remplacée par K variables binaires. B- Table de Burt. C'est une matrice carrée symétrique contenant tous les tableaux de contingence des variables prises deux à deux

3.4.2. Enrichissement – Loi hypergéométrique

La loi hypergéométrique est une loi de probabilité discrète très proche de la loi binomiale. Cette loi tente d'estimer si pour un groupe de R composés parmi lesquels r partagent des caractéristiques communes (pathologies) sur un ensemble de N composés dont n sont associés à une pathologie, le ratio $\frac{r}{R}$ est plus grand que le ratio $\frac{n}{N}$ (Figure 27).

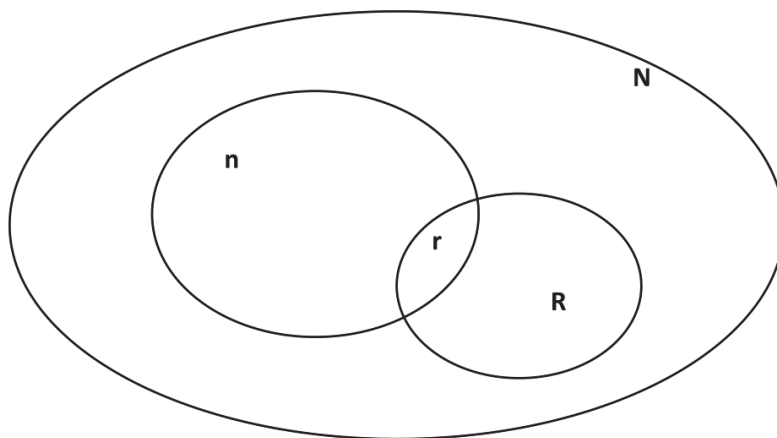


Figure 27. Description des ensembles de la loi hypergéométrique

Ensemble de la loi hypergéométrique. N est le nombre total de composés. R le nombre de composés dans le groupe d'intérêt. n est le nombre de composés associés à la pathologie. r est le nombre de composés associés à la pathologie dans le groupe d'intérêt.

La probabilité permettant de déterminer si une pathologie est sur-représentée dans un groupe par rapport à l'ensemble total des pathologies est obtenue à partir d'un calcul de z-score. Ce dernier est calculé par la soustraction du nombre observé de composés associés à la pathologie d'intérêt par le nombre attendu de composés divisé par la déviation standard du nombre de composés attendus :

$$zscore = \frac{\left(r - n \frac{R}{N}\right)}{\sqrt{n \left(\frac{R}{N}\right) \left(1 - \frac{R}{N}\right) \left(1 - \frac{n-1}{N-1}\right)}}$$

Un z-score positif reflète l'enrichissement de la pathologie dans le groupe R par rapport à la population N . Cependant cet enrichissement n'est pas toujours statistiquement significatif et c'est pour cela qu'un seuil à partir duquel le z-score reflètera un enrichissement statistique doit être défini. Ce dernier est utilisé pour déterminer une pvalue qui reflète, elle aussi, l'enrichissement d'un groupe. Un z-score haut et une faible pvalue correspondent à un enrichissement.

3.4.3. Méthodes de prédiction de classe

3.4.3.1. Partial least Square regression.

La *Partial Least Square regression* (PLS) ou régression par les moindres carrés est une version supervisée de l'ACP. Le concept de la régression PLS est de créer à partir d'un tableau de n observations décrites par p variables, un ensemble de h composantes où $h < p$. À la différence de l'ACP, cette méthode s'adapte parfaitement bien lors de l'absence de données. La détermination du nombre de composantes à retenir est en général fondée sur un critère mettant en jeu une validation croisée ou bien est fixée manuellement.

La PLS considère deux types de variables : la ou les variables dépendantes (notées Y_i) (pour une variable on parle de régression PLS1, pour plusieurs de régression PLS2) dont la variance est expliquée par un nombre de variables indépendantes (notées X_i). L'algorithme suit trois étapes successives pour réaliser la régression. La première est la recherche des m composantes principales orthogonales $t_h = Xa_h$ explicatives de leur propre groupe et bien corrélées à Y . Pour cela, et comme pour l'ACP, on cherche les composantes principales $t_h = Xa_h$ maximisant le critère de covariance $Cov(Xa_h, y)$ sous des contraintes d'orthogonalités entre t_h et les t_{h-1} autres composantes. Ensuite, le calcul de chaque composante est effectué. Enfin l'expression de la régression en fonction de X est faite.

La PLS projette ensuite les variables indépendantes sur des composantes principales, qui à la différence de l'ACP, sont orientées en fonction de leurs relations avec les variables. Le système ainsi obtenu peut-être analysé comme pour la PCA à l'aide des matrices de coordonnées ou de poids. Toutefois, la PLS apporte une information supplémentaire : l'importance de chaque variable dans le modèle. Cette importance est traduite par un indice noté VIP pour *Variable Importance in the Projection* permettant de pondérer chaque variable afin de définir un modèle de prédiction. Une variable sera considérée comme importante pour la prédiction si son indice $VIP > 0.8$.

3.4.3.2. Technique d'apprentissage supervisé – les séparateurs à vastes marges

Le séparateur à vastes marges ou *Support Vector Machines* (SVM) est une méthode d'apprentissage supervisé (Vapnik 1998) très utilisée dans le domaine de la bio-informatique (Noble 2006; Yang 2004). Ce type d'apprentissage sous-entend que la méthode utilise un jeu d'apprentissage pour la création d'un modèle. Chaque classe d'individus présents dans ce jeu d'apprentissage doit être connue. Le principe du SVM est de représenter les données par des vecteurs et à projeter ces derniers dans un espace à plusieurs dimensions afin de déterminer si un individu choisi appartient ou non à la

classe d'intérêt. Pour cela, et grâce à une étape d'apprentissage, le SVM détermine un hyperplan de séparation entre les vecteurs appartenant (exemples) à la classe et ceux qui n'appartiennent pas à celle-ci (contre-exemples). Il ne reste ainsi plus qu'à visualiser la position de l'individu sélectionné par rapport à cet hyperplan pour déterminer son appartenance au groupe (Figure 28).

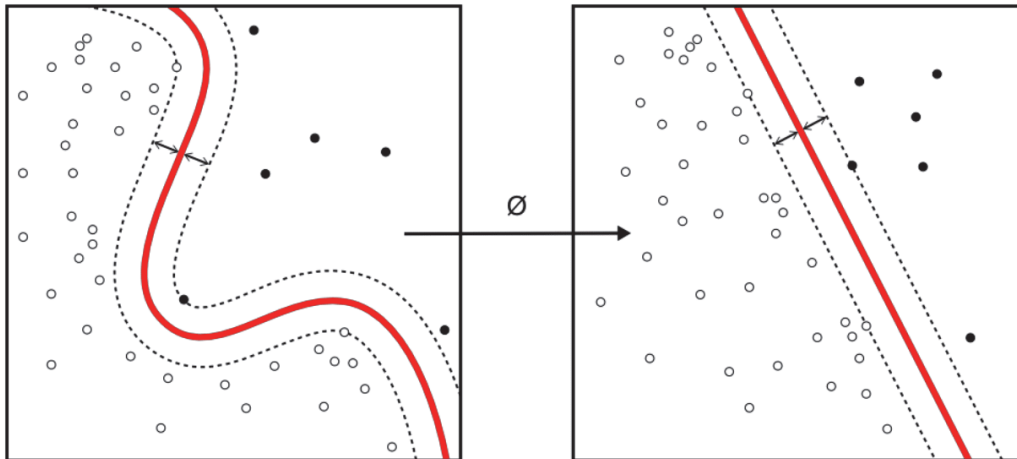


Figure 28. Discrimination des données par SVM

La méthode SVM permet de discriminer les valeurs par la création d'un hyperplan de séparation (ligne rouge). Le SVM maximise la distance entre l'hyperplan et les exemples/contre-exemples les plus proches (vecteurs supports). Cet espace entre l'hyperplan et les vecteurs supports permet de minimiser les erreurs de classifications lors de la classification de données inconnues.

La force du SVM repose dans sa capacité à maximiser la distance entre l'hyperplan et les exemples/contre-exemples les plus proches (vecteurs supports). Cet espace entre l'hyperplan et les vecteurs supports permet de minimiser les erreurs de classification des données inconnues. Afin de classer les données de façon discriminante, le SVM utilise une fonction de décision définie par :

$$f : \chi \rightarrow \{-1, +1\}$$

Cette fonction associe à chaque donnée χ une classe. La donnée sera un exemple si sa valeur est égale à +1 ou sera un contre-exemple si la valeur est égale à -1. À partir de là, il est possible de définir le risque réel de mauvaise classification. Ce risque est défini par l'équation suivante :

$$R(f) = \int_{\chi U} L[f(x), u] dP(x, u)$$

Le risque empirique estime le nombre d'erreurs de classification dans le jeu de données d'apprentissage. C'est pourquoi il est absolument nécessaire de connaître la répartition des données de ce jeu et que ce dernier soit le plus représentatif possible des données réelles. La méthode de SVM va donc estimer la meilleure fonction minimisant le risque empirique. Cependant, si la fonction est trop simple, les risques de mauvaise classification du jeu d'apprentissage et donc le risque empirique seront trop grands. De même si la fonction cherche à trop bien représenter le jeu de données, il existe un risque de « surapprentissage » tendant à mal classer les données. C'est pourquoi il est nécessaire d'assurer un compromis entre efficacité et généralité lors de la classification du jeu d'apprentissage. Pour pallier à ce problème, il est possible de calculer un indicateur de complexité de la fonction de décision.

Une fois le modèle SVM mis en place grâce au jeu d'apprentissage, il est possible d'utiliser le jeu de données à classer. Comme pour le jeu de données d'apprentissage, les objets à classer sont vectorisés et projetés dans un espace à plusieurs dimensions définies par le jeu d'apprentissage. La détermination de la classe d'un objet est faite en fonction de sa position par rapport à l'hyperplan de séparation.

3.5. Méthodes appliquées dans le cadre du projet ChemPSy

Durant ma thèse, deux ensembles de données ont été utilisés dans l'élaboration du projet ChemPSy. Initialement, le projet ChemPSy était articulé autour des données qualitatives disponibles de la CTD. Néanmoins, la nette amélioration des résultats sur la base de données quantitative nous a fortement incité à assembler un autre jeu de données à partir de GEO. Ainsi, si les méthodes appliquées sur chacun de ces jeux de données sont similaires, le prétraitement des informations issues de la CTD et de GEO est radicalement différent. C'est pourquoi je séparerai cette partie en deux : La première décrivant les méthodes appliquées aux données de la CTD et une deuxième pour les données issues de GEO.

3.5.1. Méthodes appliquées sur les données de la CTD

3.5.1.1. Constitution du jeu de données transcriptomique

Les données initialement utilisées dans ChemPSy proviennent de la CTD (cf. 3.2.5 Bases de données). L'extraction des données de la CTD et leur mise à en forme ont été effectuées via un script Tcl/Tk. Trois fichiers ont ainsi été téléchargés et formatés. Le fichier 'chemicals-diseases' décrit les relations publiées entre les composés chimiques et des pathologies ou phénotypes délétères. De la même manière, le fichier 'chemicals-genes' centralise tous les effets sur l'expression des gènes des substances chimiques décrits dans la littérature. Enfin, le fichier 'genes-disease' décrit les associations publiées entre les gènes et les pathologies ou phénotypes délétères.

À partir de ces informations, quatre matrices ont été constituées: *pos.predictive.data* et *neg.predictive.data* (lignes = composés, colonnes = gènes) dont les cellules correspondent aux nombres d'articles scientifiques qui référencent une induction (*pos.predictive.data*) ou une répression (*neg.predictive.data*) de l'expression d'un gène donné (colonne) par un composé donné (ligne); *pos.topredict.data* et *neg.topredict.data* dont les cellules correspondent au nombre d'articles scientifiques référençant une association positive (induction de la pathologie) ou une association négative (traitement de la pathologie) entre une pathologie humaine donnée (colonne) et un composé donné (ligne).

De nouvelles matrices, nommées *sum*, ont été créées par la soustraction des matrices pos et neg (ex. *pos.predictive.data* - *neg.predictive.data* = *sum.predictive.data*). Cette nouvelle matrice permet d'évaluer l'action d'un composé sur l'expression d'un gène (ou d'un phénotype) d'après la littérature. Enfin, la dernière étape de transformation correspond à une simplification des matrices *sum* en une

matrice dite « binaire » : *bin.predictive.data* et *bin.topredict.data*. De ce fait, toutes les valeurs inférieures à 0 prendront une valeur de « -1 » ; les valeurs égales à 0 prendront une valeur de « 0 » et celles supérieures à 0 prendront une valeur de « 1 » (Figure 29).

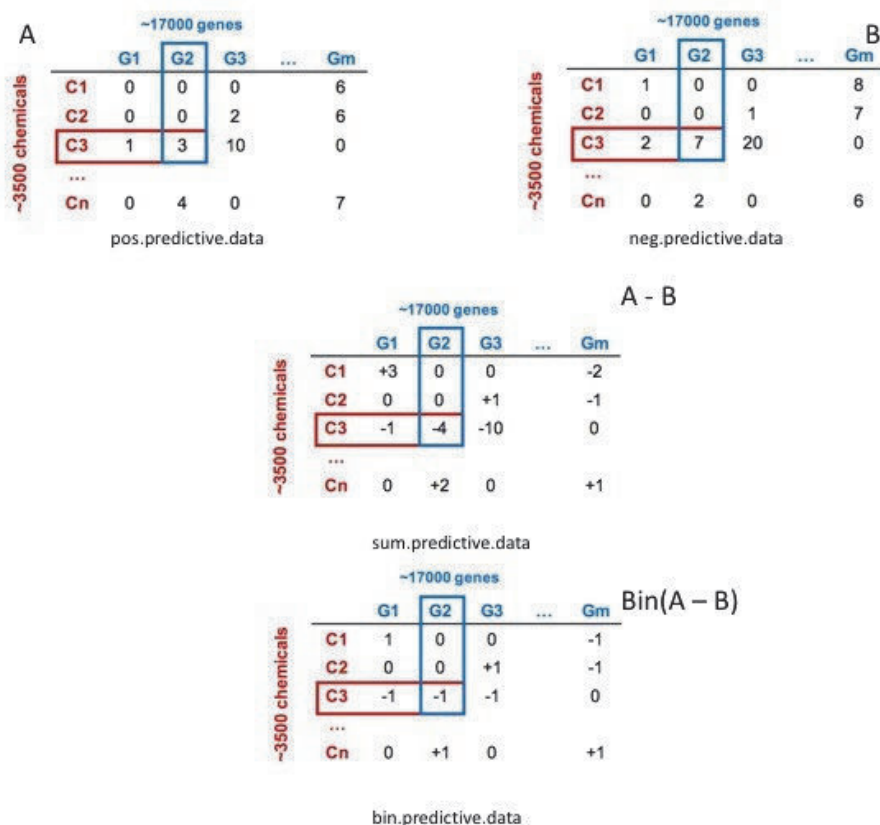


Figure 29. Travail effectué sur les données obtenues à partir de la CTD

Les matrices *pos.predictive.data* (A) et *neg.predictive.data* (B) sont les matrices dont chaque cellule correspond au nombre d'articles scientifiques référant une induction (*pos.predictive.data*) ou une répression (*neg.predictive.data*) de l'expression d'un gène donné par un composé donné. La matrice, *sum*, est obtenue par la soustraction des matrices *pos* et *neg* ($A - B = \text{sum.predictive.data}$). Les données de la matrice *sum* sont simplifiées de façon à ce que toutes les valeurs inférieures à 0 prennent une valeur de « -1 » ; les valeurs égales à 0 prennent une valeur de « 0 » et celles supérieures à 0 prennent une valeur de « 1 ».

3.5.1.2. Analyses multivariées et classification

L'exploitation des données de la CTD a été réalisée par des analyses multivariées en utilisant le package FactoMineR. À partir de la matrice *bin.predictive.data* obtenue, des ACM ont été appliquées afin de sélectionner les composantes les plus informatives, c'est-à-dire portant au moins 80% de l'information cumulée. Les coordonnées des individus (substances chimiques) sur les composantes

sélectionnées sont ensuite utilisées pour les classer sur la base de leur signature toxicogénomique, les gènes étant utilisés comme des variables. Pour cela l'outil de classification du package FactoMineR, HCPC ou Classification Hiérarchique sur Composantes Principales a été utilisé.

L'HCPC repose sur une méthode de classification ascendante hiérarchique (CAH) classique. Il s'agit d'une technique statistique visant à partitionner une population en différentes classes ou sous-groupes. Cette méthode regroupe au sein d'une même classe les individus présentant le plus de similarités possible afin d'obtenir une inertie intraclasse (homogénéité des distances) la plus faible possible ainsi qu'une inertie interclasse la plus grande possible (centre de gravité de chaque classe le plus éloigné possible).

Le principe de la CAH est de rassembler des individus selon un critère de ressemblance défini au préalable se présentant sous la forme d'une matrice de distances, exprimant la distance existant entre chaque individu pris deux à deux. Deux observations identiques auront une distance nulle. Plus les deux observations seront dissemblables, plus la distance sera importante. La CAH va ensuite rassembler les individus de manière itérative afin de produire un dendrogramme ou arbre de classification. Cette agrégation est obtenue par l'utilisation de différentes méthodes, les plus classiques étant la méthode de Ward, le saut minimum et le saut maximum. La classification est ascendante, car elle part des observations individuelles ; elle est hiérarchique, car elle produit des classes ou groupes de plus en plus vastes, incluant des sous-groupes en leur sein. En découpant cet arbre à une certaine hauteur choisie, on produira la partition désirée. Cette découpe se fait en prenant en compte la perte d'inertie interclasse lors du passage d'un système à n groupes à un système à $n + x$ groupes. Le dendrogramme est alors coupé avant la forte perte d'inertie du système.

3.5.2. Méthodes appliquées sur les données issues de GEO

3.5.2.1. Constitution du jeu de données transcriptomique

L'objectif de cette étape est de constituer un jeu de données quantitatives massif de toxicogénomique basé sur des approches transcriptomiques. Pour cela, le dépôt de données GEO (cf. 3.2.5. Bases de données) a été utilisé.

Pour accumuler les signatures toxicogénomiques d'un maximum de substances chimiques, je me suis focalisé sur une espèce en particulier et une technologie. Le rat a été choisi, car il s'agit d'une espèce modèle communément utilisée en laboratoire notamment en toxicologie pour évaluer la toxicité des substances chimiques. La technologie choisie est celle des puces à ADN de type Affymetrix Rat 230 2.0,

très utilisée dans des projets majeurs de toxicogénomique et permettant d'interroger l'expression de plus de 13877 gènes. J'ai donc parcouru la banque GEO avec minutie afin de sélectionner les études les plus appropriées effectuées en toxicogénomique utilisant le modèle rat et cette technologie.

À partir de ces recherches, 18 études ont été sélectionnées. À la suite du téléchargement des fichiers bruts (format CEL), une uniformisation des informations associées à chaque étude a été effectuée afin de pouvoir les rendre comparables entre elles. Par exemple, les unités de temps, les doses, le nom du tissu et de l'organisme ont été uniformisés. Cette étape a également permis de standardiser le nom des conditions expérimentales (CE) et de les attribuer à chaque fichier CEL. Une CE se définit ici comme la combinaison d'une substance, à une dose donnée, à un temps d'exposition donné, dans un tissu donné, dans une espèce donnée, dans une étude donnée [CE = f (Substance, Dose, Temps, Tissu, Espèce, Étude)].

3.5.2.2. Prétraitement des données et filtration statistique

Les fichiers CEL associés à une CE donnée ainsi qu'à la condition contrôle correspondante sont normalisés afin de permettre la comparaison des données de transcriptomique et l'identification des signatures toxicogénomiques de chaque CE.

Dans un premier temps, la qualité individuelle des fichiers bruts d'expression (fichiers CEL) a été vérifiée manuellement. L'image de la puce est reconstituée sous R pour chaque puce (chaque fichier CEL) afin de détecter de potentiels problèmes d'hybridation. Si celles-ci représentent plus de 20% de la surface de la puce, cet échantillon est retiré du jeu de données (Figure 30).

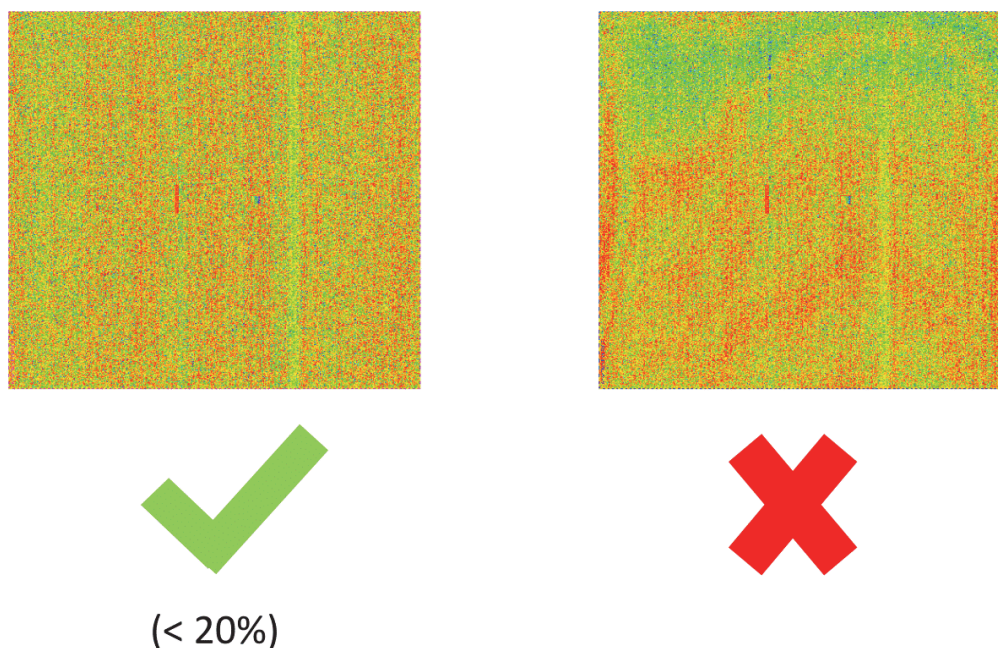


Figure 30. Représentation d'une puce à ADN

Représentation visuelle d'une puce à ADN. Celle-ci permet de détecter de potentielles erreurs d'hybridation. Si celles-ci représentent plus de 20% de la surface de la puce, cet échantillon est retiré du jeu de données. À gauche, un exemple de puce conservé dans notre jeu de donnée et à droite un exemple de puce retirée du jeu.

Les fichiers CEL restants sont ensuite normalisés par la méthode RMA (cf. 3.1.1.2 Environnement R). Au cours de cette étape, j'ai notamment utilisé l'environnement CDF de la puce Affymetrix Rat 230 2.0 fourni par BrainArray (Dai et al. 2005) afin d'obtenir directement une valeur d'expression unique pour chaque gène (Entrez gene ID) simplifiant ainsi grandement les étapes de conversion et de comparaison des résultats.

Une fois les données normalisées, les échantillons exposés ont ensuite été comparés aux échantillons contrôles afin d'identifier la signature toxicogénomique correspondant à chaque CE, c'est-à-dire l'ensemble des gènes dont l'expression est significativement altérée suite à une exposition à un composé chimique. Pour cela, nous avons à la fois utilisé un seuil sur le ratio d'expression (ou fold-change) entre la moyenne des échantillons exposés et la moyenne des échantillons contrôles, mais également un test statistique avec modèle linéaire d'estimation de la variance, et correction pour les tests multiples *via* la librairie Limma (cf. 3.1.1.2. Environnement R). Brièvement, les gènes différenciellement exprimés présentaient un $\log_2(\text{fold-change}) > 0.585$ ou < -0.585 (correspondant à un ratio de 1.5) et une $p\text{-value} < 0.05$ (ajustée par la méthode de Benjamini-Hochberg).

Ensuite, toutes les signatures ont été rassemblées au sein de la même matrice avec en lignes les gènes interrogés et colonne les CE. Celle-ci est déclinée en deux exemplaires. La matrice log2fold-change (log2FC) qui stocke pour chaque condition la valeur du *fold-change* du gène interrogé et une matrice discrétisée composé de 0 quand la valeur du log2FC = 0, 1 quand le log2FC > 0 et -1 quand le log2FC < -1. Afin de réduire cette matrice et de garder uniquement les gènes et les CEs présentant des modifications, les gènes étant altérés dans aucune CEs ont été retirés. De même, les CEs affectant l'expression de moins de 10 gènes ont été supprimés (Réduction des matrices).

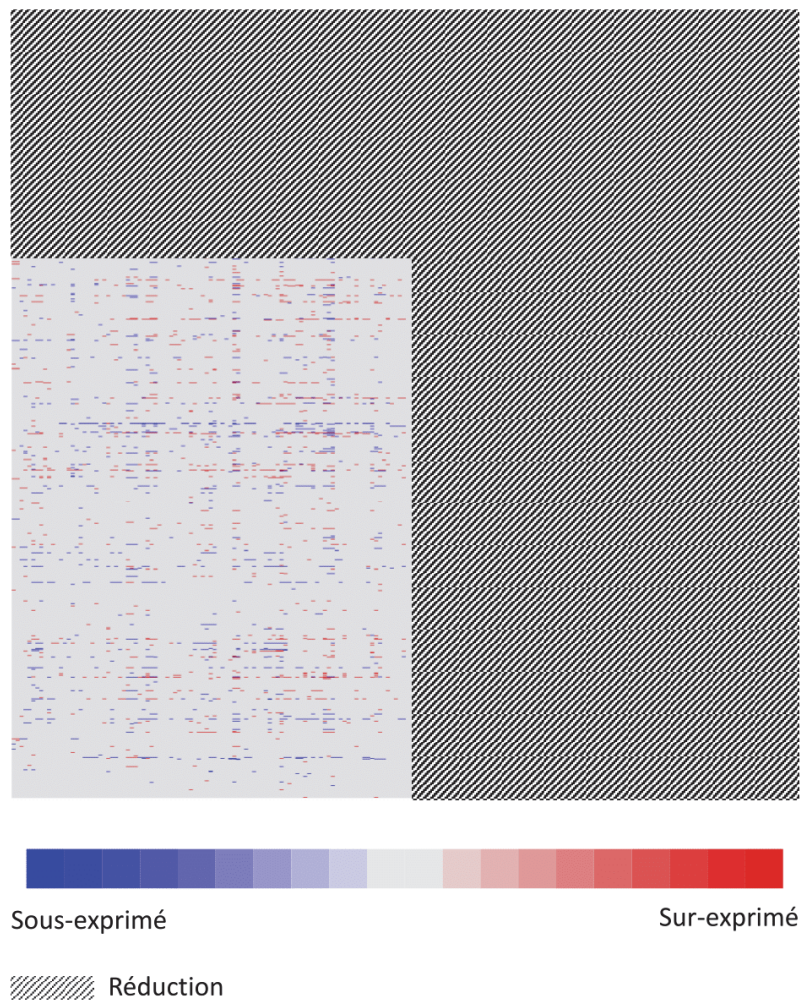


Figure 31. Réduction des matrices

Les matrices obtenues ont été réduites afin de garder uniquement les gènes et les CEs présentant des modifications. Les gènes étant sur ou sous exprimés dans moins d'une CE ont été retirés. De même, les CEs affectant l'expression de moins de 10 gènes ont été supprimés.

Une analyse exploratoire a ensuite été effectuée sur ces données. Cette dernière consiste à appliquer sur les deux matrices une ACP où les individus correspondent aux CE et les variables aux gènes. L'objectif de cette étape est de réduire les matrices et ne garder uniquement les composantes expliquant plus de 80% de la variabilité du système.

Au final, quatre matrices seront utilisées pour la suite de l'analyse : les matrices dites « brutes », c'est-à-dire la matrice log2FC et la matrice discrétisée ainsi que les matrices issues de l'ACP, ACP-log2FC et ACP-discrétisée.

3.5.2.3. Choix des estimateurs de similitude

Afin d'estimer les distances regroupant au mieux les CE en fonction de leurs signatures toxicogénomiques ainsi que leurs effets toxicologiques, j'ai tenté de déterminer s'il existait une corrélation linéaire entre la distance toxicogénomique et la distance toxicologique. On définit ici une distance toxicogénomique la similarité évaluée par un estimateur de distance entre deux CE sur la base de leur signature toxicogénomique. Pour cette distance, les estimateurs utilisés sont : la distance euclidienne, la corrélation de Pearson, le Hubert's score, l'indice de Kappa et l'index de Jaccard. La distance toxicologique correspond à la proximité entre deux CE sur la base de l'annotation des composés chimiques fournie par la CTD. Celle-ci est estimée selon l'index de Jaccard, le coefficient de Kappa, l'Hubert score, le zscore, l'odds ratio et la concordance.

La corrélation entre les distances toxicogénomiques et les distances toxicologiques a ensuite été calculée afin de sélectionner les deux meilleurs estimateurs toxicogénomique, c'est-à-dire le couple d'estimateurs de distance maximisant la corrélation entre les distances toxicogénomiques et toxicologiques.

3.5.2.4. Discrimination de classes de composés chimiques

L'objectif ici est de classifier les CE en fonction de leurs signatures toxicogénomiques *via* différentes méthodes de classification non supervisées. Comme pour les distances, les meilleures techniques de classification seront sélectionnées pour la suite du projet. Afin de regrouper les CE présentant une signature toxicogénomique proche, trois méthodes de classification non-supervisées distinctes (Mclust, Hopach, Dynamic tree cut) (cf. 3.1.1.2 Environnement R) utilisant les mesures de distance préalablement sélectionnées (euclidienne, log de l'euclidienne et corrélation) ont été utilisées.

Si la classification obtenue n'est pas satisfaisante, c'est-à-dire si le nombre de groupes constitués par une de ces méthodes est inférieur à 50, une méthode récursive est employée. Cette méthode consiste à réappliquer la méthode de clustering pour chaque cluster contenant plus de 50 éléments. En d'autres termes, si un cluster de 75 CEs est obtenu avec la méthode Mclust, Mclust est réutilisé sur ce cluster afin de le sous-diviser en sous-groupes.

3.5.2.5. Association des classes avec des effets toxicologiques

Dans le but d'évaluer la pertinence des classes obtenues, une méthode d'association entre les groupes de composés et les pathologies humaines a été développée. Cette étape utilise le fichier 'chemicals-diseases' de la CTD décrit ci-dessus (cf 3.5.1 Méthodes appliquées sur les données de la CTD). Pour cela, les effets toxiques les plus sur-représentés dans les différents groupes sont évalués à l'aide de la loi hypergéométrique. Dans notre cas, un seuil de $p\text{-value} < 0,005$ a été utilisé. Enfin, pour considérer qu'un groupe est enrichi les ensembles n et r doivent contenir un nombre minimum de CEs fixé à 3.

Un enrichissement est calculé afin de déterminer quelles sont les pathologies sur-représentées sur la base des composés présents dans le groupe (utilisation du fichier 'chemicals-diseases' de la CTD). Parmi les pathologies présentes dans la CTD, le terme « perturbateur endocrinien » n'existait pas. Pour pallier à ce problème, j'ai effectué un état de la littérature sur les composés présents dans notre ensemble de données afin de discriminer les substances possédant des propriétés de PE potentiels. Les effets toxiques les plus sur-représentés dans les différents groupes sont évalués à l'aide de la loi hypergéométrique.

Parallèlement, la signature toxicogénomique de chaque classe est déterminée comme étant l'ensemble des gènes sur/sous exprimés dans au moins 80% des CEs incluses dans cette classe.

3.5.2.6. Prédiction

Cette étape cherche à définir, sur la base de la signature toxicogénomique d'un nouveau composé, à quelle classe il se rapproche le plus. Pour cela la méthode d'évaluation de modèle statistique de classification, la cross validation (cf. 3.3.2 Évaluation d'un modèle statistique de classification) a été utilisée. La cross validation s'appuie sur la reclassification d'une CE en fonction soit (i) de la méthode de SVM (cf. 3.4.3.2 Technique d'apprentissage supervisé – les séparateurs à vastes marges) ; (ii) de la PLS (cf. 3.4.3.1 Partial least Square regression) ; (iii) de sa proximité avec le centroïde des classes (moyenne des positions de chaque individu de la classe) ; (iv) sa proximité avec le médoïde des classes (médiane des positions de chaque individu de la classe) ; (v) de sa proximité avec la position d'une autre CE. Dans ce cas la CE à prédire est classée dans le même groupe que la CE la plus proche.

Pour chaque méthode, 20% du jeu de données a été utilisé comme jeu de « test » et une technique de rééchantillonnage bootstrap a été appliquée afin de valider la robustesse des reclassifications (une centaine d'itérations). La sensibilité et la spécificité moyenne ont été calculées pour chaque rééchantillonnage ainsi que le taux d'erreur pour les reclassifications basées sur le centroïde, médoïde et proximité de CE.

4. Résultats et discussion

4.1. ChemPSy : un système de priorisation pour les substances chimiques

4.1.1. Résultats obtenus à partir de la CTD

4.1.1.1. Création du jeu de données

La première étape de ce projet a consisté en l'intégration des données de la CTD. Grâce à cela, il est possible d'interroger l'ensemble des associations connues (avec des publications scientifiques à l'appui) entre les composés chimiques référencés dans cette banque et les altérations positives ou négatives de l'expression des gènes humains qu'ils induisent ou les pathologies humaines (composés associés à des pathologies humaines ou phénotypes délétères) parmi lesquelles des pathologies et désordres reproductifs et de la perturbation endocrine. Les matrices *pos.predictive.data* et *neg.predictive.data* mettent en relation les 1477 composés de la CTD avec 19360 gènes avec en moyenne 140 gènes activés et 111 réprimés associés à un composé. Les matrices *pos.topredict.data* et *neg.topredict.data* associent, quant à elles, ces 1477 substances chimiques à 3336 phénotypes différents avec en moyenne 48 phénotypes induits par composé et 28 phénotypes traités/soignés par composé. La matrice *sum.predictive.data* regroupe 19351 gènes pour 1477 produits chimiques et présente une moyenne de 30 gènes exprimés (induits - réprimés) par composé. Enfin, la matrice *sum.topredict.data* relie les 1477 composés aux 3336 phénotypes avec en moyenne 22 phénotypes exprimés (induits – traité) par substance. Bien qu'informatives sur la tendance d'un composé, les matrices *sum* présentent un souci d'interprétation de la valeur 0. En effet, cette valeur correspond soit à l'absence d'association entre le composé et le gène/pathologie, soit à la soustraction d'un nombre équivalent d'interactions négatives et positives. Il n'est donc pas possible de définir si le 0 correspond à une absence d'informations ou bien à la concaténation d'informations contraires.

4.1.1.2. Classification des données

Les 1477 composés ont été regroupés dans 111 groupes à la suite d'une classification hiérarchique précédée d'une ACM. Pour la grande majorité d'entre elles, ces classes sont constituées d'une ou deux substances, alors qu'un groupe particulier contient à lui seul plus de 1200 composés sur les 1477 au total (Figure 32).

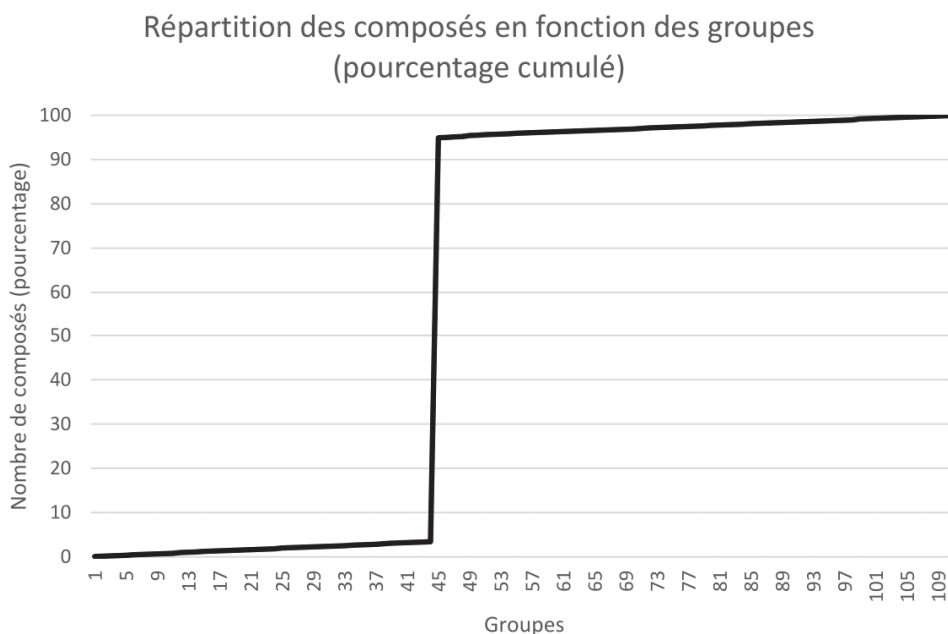


Figure 32. Répartition des composés en fonction des 111 groupes en pourcentage cumulé

Les 1477 composés ont été regroupés dans 111 classes. Cependant, la majorité d'entre elles sont constituées d'une ou deux substances, alors qu'un groupe particulier (45) contient à lui seul plus de 1352 composés sur les 1477 au total soit 91% de nombre de substance. Ceci se traduit par l'augmentation soudaine de la courbe.

Face à l'impossibilité d'obtenir une répartition homogène des composés, d'autres techniques de classification ont été utilisées (Mclust, Mclust récursif et Dynamic Tree Cut). Les résultats obtenus recoupent le précédent résultat. La méthode Mclust répartit les substances en 9 groupes dont un groupe contenant 874 composés soit 59% du jeu de donnée. La méthode récursive de Mclust répartit les composés dans 46 groupes dont 41 constitués de moins de 5 composés et 5 de plus de 200. Dynamic Tree Cut sépare les composés en 2 groupes : un de 1350 composés et un deuxième de 127.

Au vu des résultats sur les différentes techniques de classification, les stratégies utilisées ne m'ont donc pas permis de regrouper les substances analysées en classes exploitables à partir des données de la CTD. Une explication plausible à cela serait que les signatures toxicogénomiques de la CTD

manquent de fiabilité, car les données à partir desquelles elles sont extraites sont (i) qualitatives plutôt que quantitatives puisqu'elles représentent un nombre d'occurrences dans la littérature, (ii) incomplètes et hétérogènes car chaque composé a été étudié avec un protocole différent, certains à haut débit (puces à ADN) et d'autres de manière ciblée (PCR quantitative par exemple). Enfin (iii) les données manquantes ne peuvent être différenciées des absences de variation d'expression. Nous avons alors décidé de constituer un nouveau jeu de données, basé sur les données brutes d'expériences de puces à ADN, qui permettrait de pallier aux différents problèmes rencontrés avec la CTD.

4.1.2. Résultats obtenus à partir de GEO

4.1.2.1. Jeu de données

Suite à l'exploration du site de dépôt de données, GEO je me suis orienté vers la technologie des puces à ADN du type « Affymetrix GeneChip Rat 230 2.0 », car celle-ci a été très largement utilisée dans des expériences de toxicogénomique. Cette plateforme permet d'interroger simultanément l'expression de 13877 gènes chez le rat. Dans son ensemble, le jeu de données obtenu après téléchargement est composé de 18 études contenant des informations sur la réponse transcriptionnelle de 7 tissus de rat (foie, rein, cœur, muscle lisse et dans une moindre mesure cerveau, testicule et ovaire) à 452 composés chimiques, pour un total de 7465 conditions expérimentales (CEs), correspondant à plus de 26357 échantillons, et donc autant de fichiers bruts (fichiers CEL de puces Affymetrix GeneChip Rat 230 2.0) (Tableau IX).

GSE	PMID	Tissu	Conditions expérimentales	Nombre de composé
GSE10748	19212759	Cerveau	5	1
GSE57800	25058030	Cœur	206	88
GSE19366	21865292	Rein	24	3
GSE57811	25058030, 26260164	Rein	365	139
TG-GATEs	25313160	Rein	1363	150
GSE57815	25058030	Foie	654	200
TG-GATEs	25313160	Foie	4776	150
GSE8238	17127748	Ovaire	1	1
GSE10557	17660231	Ovaire	1	1
GSE32890	/	Ovaire	1	1
GSE9480	/	Testicule	2	1
GSE10412	19423681	Testicule	6	3
GSE10919	18042343	Testicule	3	1
GSE20245	20566332	Testicule	5	2
GSE20952	/	Testicule	10	10
GSE25196	21266533	Testicule	1	1
GSE57816	25058030	Muscle	42	21

Tableau IX. Résumé des données intégrées dans ChemPSy

Parmi ces 18 études, deux d'entre elles composent la grande majorité des informations :

- DrugMatrix (Ganter et al. 2006) est un projet du National Institute of Environmental Sciences (NIES, USA) qui a déterminé la réponse transcriptionnelle à 376 composés différents (à plusieurs doses, à plusieurs temps d'exposition, en triplicat) dans 5 tissus chez le rat.

- Open TG-GATEs (Toxicogenomics Project-Genomics Assisted Toxicity Evaluation system) (Igarashi et al. 2015) est le fruit de la collaboration du NIB (National Institute of Biomedical Innovation), du NIES et d'une quinzaine de compagnies pharmaceutiques. Il répertorie les réponses transcriptionnelles à 150 composés (à plusieurs doses, à plusieurs temps d'exposition, en duplicat) dans le rein et le foie chez le rat.

4.1.2.2. Filtration statistique

La normalisation (RMA) et la filtration statistique (fold-change ≥ 1.5 et test statistique Limma, p -value < 0.05 ajustée par la méthode de Benjamini-Hochberg) des données ont permis d'obtenir une signature toxicogénomique pour 7076 CE. Deux matrices de 7076 lignes (une par CE) et 13877 colonnes (une par gène interrogeable) ont alors été constituées. Dans la première matrice, la valeur du $\log_2(\text{fold-change})$ a été reportée pour chaque gène dans chaque condition expérimentale. Dans la seconde, les données ont été discrétisées. Pour une CE donnée, un gène peut avoir 3 attributs : +1 (le gène est surexprimé suite à l'exposition, $\log_2(\text{fold-change}) > 0.585$) ; -1 (le gène est sous-exprimé suite à l'exposition, $\log_2(\text{fold-change}) < -0.585$) ; 0 (le gène n'est pas différenciellement exprimé, $-0.585 < \log_2(\text{fold-change}) < 0.585$). Par la suite, ces deux matrices ont été réduites en ne conservant que les conditions expérimentales pour lesquelles la signature toxicogénomique contenait au moins 10 gènes dérégulés. De la même manière, seuls les gènes présentant une variation significative dans au moins une CE ont été conservés. Ces deux matrices finales sont donc composées de 3022 lignes (CEs) et 11434 colonnes (gènes). Les 3022 CE correspondent à 410 composés chimiques distincts. Le foie (2246 CE) et le rein (573 CE) y sont les organes les plus représentés. La majorité des composés ne sont représentés que par une seule CE tandis que d'autres sont représentés par plusieurs dizaines de CE : 20 pour la gentamicine (antibiotique), 29 pour le cis-platine (chimiothérapie) et 56 pour l'éthinylestradiol (œstrogène actif par voie orale le plus utilisé au monde, notamment dans la plupart des pilules contraceptives combinées).

4.1.2.3. Sélection des meilleurs estimateurs de distance

Afin de vérifier s'il existe une corrélation linéaire entre la distance toxicogénomique (signatures géniques) et la distance toxicologique (ontologie des composés chimiques), j'ai comparé plusieurs

estimateurs de distance. L'objectif ici étant de sélectionner les deux meilleures distances à utiliser pour les solutions de classification afin de diminuer le nombre de méthodes à appliquer.

	CC	DE	CC (T)	Log(DE)	JI	Ratio	HG	Z	Corr	Kap	HG (T)	Z (T)	JI (T)	Ratio (T)
PCABIN - Corr	0,24	0,20	0,19	0,19	0,15	0,15	0,14	0,14	0,14	0,12	0,06	0,06	0,07	0,07
PCALOG2FC - Corr	0,23	0,19	0,19	0,18	0,14	0,14	0,14	0,14	0,14	0,12	0,07	0,07	0,07	0,07
BIN - HG	0,16	0,12	0,15	0,11	0,13	0,13	0,14	0,14	0,14	0,12	0,10	0,10	0,09	0,09
BIN - HG (T)	0,16	0,11	0,15	0,10	0,13	0,13	0,14	0,14	0,14	0,12	0,10	0,10	0,09	0,09
BIN - Z	0,16	0,12	0,15	0,11	0,13	0,13	0,14	0,14	0,14	0,12	0,10	0,10	0,09	0,09
BIN - Z (T)	0,16	0,11	0,15	0,10	0,13	0,13	0,14	0,14	0,14	0,12	0,10	0,10	0,09	0,09
LOG2FC - Corr	0,15	0,10	0,14	0,09	0,13	0,13	0,14	0,14	0,14	0,11	0,10	0,10	0,09	0,09
BIN - Corr	0,15	0,10	0,13	0,09	0,13	0,13	0,14	0,14	0,14	0,11	0,10	0,10	0,09	0,09
BIN - JI (T)	0,15	0,10	0,14	0,09	0,12	0,12	0,14	0,14	0,14	0,11	0,10	0,10	0,09	0,09
BIN - Ratio (T)	0,15	0,10	0,14	0,09	0,12	0,12	0,14	0,14	0,14	0,11	0,10	0,10	0,09	0,09
BIN - JI (T)	0,14	0,09	0,13	0,09	0,12	0,12	0,13	0,13	0,13	0,11	0,09	0,09	0,09	0,09
BIN - Ratio	0,14	0,09	0,13	0,09	0,12	0,12	0,13	0,13	0,13	0,11	0,09	0,09	0,09	0,09
BIN - CC (T)	0,05	0,02	0,05	0,02	0,06	0,06	0,06	0,06	0,06	0,06	0,05	0,05	0,04	0,04
BIN - CC (T)	0,04	0,01	0,04	0,01	0,07	0,07	0,08	0,08	0,08	0,07	0,06	0,06	0,05	0,05
LOG2FC - DE	-0,12	-0,09	-0,09	-0,09	-0,08	-0,08	-0,08	-0,08	-0,08	-0,07	-0,01	-0,01	-0,02	-0,02
PCALOG2FC - DE	-0,12	-0,09	-0,09	-0,09	-0,08	-0,08	-0,08	-0,08	-0,08	-0,07	-0,02	-0,02	-0,02	-0,02
BIN - Kap	-0,12	-0,16	-0,10	-0,15	-0,03	-0,03	-0,03	-0,03	-0,03	-0,01	0,02	0,02	0,02	0,02
BIN - DE	-0,15	-0,13	-0,12	-0,13	-0,09	-0,09	-0,09	-0,09	-0,09	-0,07	-0,01	-0,01	-0,02	-0,02
PCABIN - DE	-0,15	-0,13	-0,12	-0,13	-0,09	-0,09	-0,09	-0,09	-0,09	-0,07	-0,01	-0,01	-0,02	-0,02
LOG2FC - Log(DE)	-0,18	-0,15	-0,14	-0,15	-0,10	-0,10	-0,10	-0,10	-0,10	-0,09	-0,03	-0,03	-0,04	-0,04
PCALOG2FC - Log(DE)	-0,18	-0,15	-0,15	-0,15	-0,11	-0,11	-0,10	-0,10	-0,10	-0,09	-0,03	-0,03	-0,04	-0,04
BIN - Log(DE)	-0,20	-0,18	-0,16	-0,17	-0,11	-0,11	-0,11	-0,11	-0,11	-0,09	-0,02	-0,02	-0,03	-0,03
PCABIN - Log(DE)	-0,20	-0,18	-0,16	-0,17	-0,11	-0,11	-0,11	-0,11	-0,11	-0,09	-0,02	-0,02	-0,03	-0,03

Tableau X. Corrélations entre les estimateurs mathématiques de similarité toxicogénomique et d'effets toxicologiques.

Plusieurs mesures de similarité (Corr : corrélation de Pearson ; HG : Hubert's gamma score ; JI : index de Jaccard ; CC : concordance ; DE : distance euclidienne ; Kap : distance de Kappa ; Z : Zscore ; Ratio : Odd ratio) ont été utilisées pour estimer la proximité entre deux composés chimiques du point de vue de leurs signatures toxicogénomiques et de leurs signatures toxicologiques. La relation linéaire entre ces mesures de proximité toxicogénomique (lignes du tableau) et toxicologique (colonnes) a ensuite été évaluée par un calcul de corrélation (intersection entre une ligne et une colonne). Concernant les signatures toxicogénomiques, les quatre matrices décrites dans le corps du texte ont été utilisées : BIN, Matrice binaire (-1,0,1) ; PCABIN, résultat de l'ACP sur la matrice binaire ; LOG2FC, Matrice log2 fold change ; PCALOG2FC, résultat de l'ACP sur la matrice log2 fold change. Les noms de colonnes et de lignes annotés « T » (pour TRUE) indique que les mesures de similarité ont été calculées sur des valeurs pondérées. L'objectif est de diminuer la contribution des gènes et des effets toxicologiques fréquemment retrouvés dans l'ensemble des signatures toxicogénomiques et toxicologiques. Ainsi, un gène dérégulé dans de nombreuses conditions expérimentales aura un poids moins important qu'un gène dérégulé dans peu de conditions expérimentales.

Ceux-ci ont montré une grande disparité dans les résultats obtenus (Tableau X), et les distances toxicogénomiques présentant la meilleure relation linéaire avec les effets toxicologiques se sont avérées être la corrélation de Pearson (corrélation proche de 1) et la distance euclidienne (corrélation minimum). Ces deux estimateurs ont donc été choisis pour les étapes ultérieures du projet. Par ailleurs, cette étape a permis de vérifier l'hypothèse générale de mon projet de thèse, à savoir que deux substances chimiques possédant des signatures toxicogénomiques proches, voire très proches, sont associées à des effets toxicologiques similaires.

4.1.2.4. Classification

Cette étape regroupe les CE en fonction de leurs signatures toxicogénomiques par l'utilisation de différentes méthodes de classification non supervisées (Mclust, Hopach, Dynamic tree cut).

La classification directe des CE, sans utilisation de la récursivité a permis d'obtenir les résultats présentés dans le Tableau XI. Une évaluation des méthodes de classification a également été réalisée : l'objectif est ici d'éviter d'obtenir, comme précédemment avec la CTD, un groupe contenant la majorité des CE et la majorité des classes restantes ne contenant qu'une poignée de CE. Pour cela les méthodes de classification ont été utilisées avec leurs paramètres de base sur les matrices « brutes » et issues de l'ACP, sans être préalablement converties en matrices de distance euclidienne et de corrélation.

A

	Mclust		Hopach		Dynamic Tree Cut	
	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée
Matrice « brute »	9	9	33	/	21	18
Matrice ACP	10	8	35	12	47	18

B

	Mclust		Hopach		Dynamic Tree Cut	
	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée
Matrice « brute »	336	336	91	/	143	377
Matrice ACP	302	377	86	251	64	167

Tableau XI. Résultat de la classification des CE sans récursivité

A – Nombres de groupes obtenus pour chaque méthode sur les quatre matrices (discrétisée, log2fc, PCA-discrétisée, PCA log2fc). B – Nombre moyen de CE par cluster pour les différentes méthodes de classification. La méthode Hopach n'est pas compatible avec la matrice discrétisée « brute », c'est pourquoi aucun cluster n'est obtenu avec cette technique.

Au final 210 clusters sont obtenus contenant en moyenne 230 CEs par cluster. Pour chacune des méthodes de classification appliquée, le nombre de clusters est toujours inférieur à 50 classes. Ce nombre de classes est trop faible pour discriminer des signatures toxicogénomiques homogènes qui permettaient de suffisamment bien prédire les effets toxicologiques. C'est pourquoi j'ai finalement utilisé ces mêmes méthodes de manière récursive, afin d'améliorer la granularité des solutions de classification (Tableau XII).

A

	recMclust		recHopach		recDynamic Tree Cut	
	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée
Matrice « brute »	100	70	88	/	60	45
Matrice ACP	115	89	92	43	117	118

B

	recMclust		recHopach		recDynamic Tree Cut	
	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée	Log2(Fc)	Discrétisée
Matrice « brute »	30	43	34	/	50	67
Matrice ACP	26	33	32	70	26	25

Tableau XII. Résultat de la classification des CEs avec récursivité

A – Nombres de groupes obtenus pour chaque méthode récursive sur les quatre matrices (discrétisée, log2fc, PCA-discrétisée, PCA log2fc). B – Nombre moyen de CEs par cluster pour les différentes méthodes de classification récursive. La méthode Hopach n'est pas compatible avec la matrice discrétisée « brute », c'est pourquoi aucun cluster n'est obtenu avec cette technique.

Au regard de ces résultats, j'ai finalement choisi les solutions de classification de la méthode DynamicTreeCut, car celle-ci offrait le meilleur rapport nombre de CE par groupe (matrice PCA-discrétisée). Cette technique de classification a ensuite été relancée sur la matrice PCA-discrétisée préalablement convertie en matrice de distance euclidienne et en matrice de corrélation. Ainsi 95 groupes sont créés à partir de la méthode DynamicTreeCut en utilisant la distance euclidienne et 104 en utilisant la corrélation.

4.1.2.5. Évaluation de la classification

Pour chaque méthode de classification, la pertinence des groupes a été évaluée en mesurant le degré de variabilité intragroupe. En effet, les méthodes que j'ai utilisées cherchent à classer toutes les CEs dans un groupe, quitte à ce que la pertinence de ce groupe soit contestable. Cela peut ainsi donner lieu à la création dans le groupe d'un point éloigné de son centre pouvant par la suite conduire à des erreurs d'interprétation ou de prédiction. Pour chaque CE, j'ai ainsi déterminé quelles sont les CEs les plus proches en termes de distance. Les courbes d'agglomération obtenues permettent alors d'estimer si la CE a été correctement classifiée ou si sa présence est au contraire due au hasard. Une condition est considérée comme bien classifiée si l'aire sous la courbe (AUC) est supérieure à 0,8 (Figure 33)

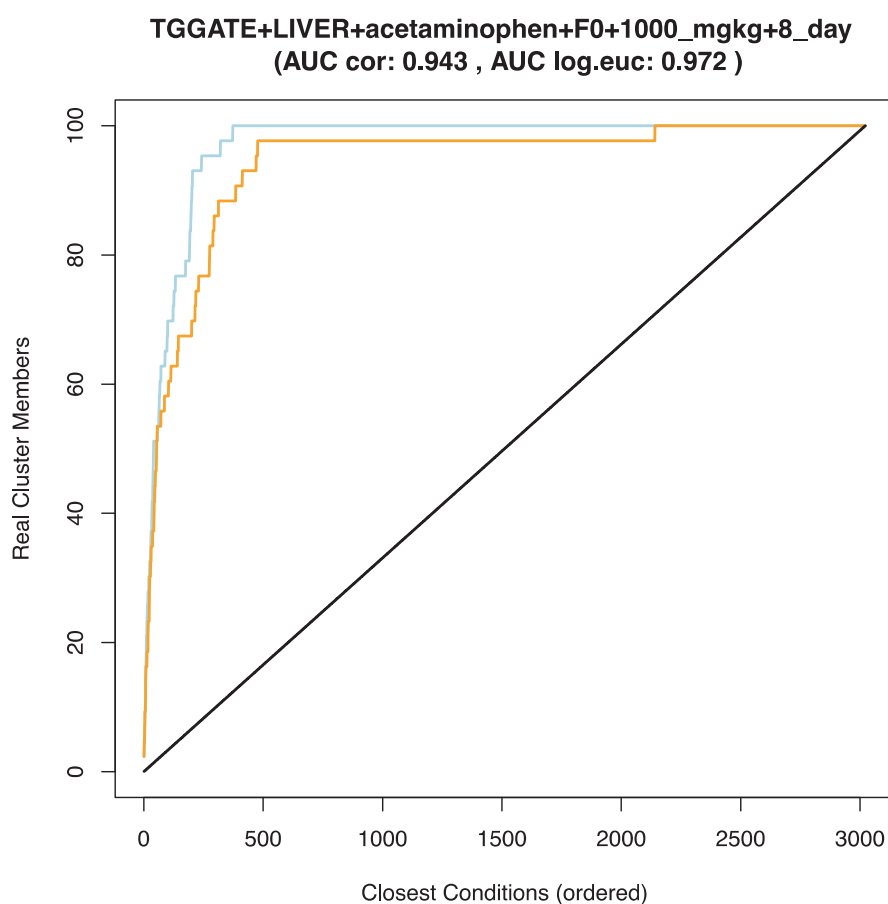


Figure 33. Exemple de courbe de classification d'une CE

Représentation de la validité de la classification d'une CE. Les CEs sont ordonnées de la plus proche à la plus éloignée de celle étudiée en utilisant la corrélation (courbe orange) et la distance euclidienne (courbe bleue). Pour chaque CE appartenant au même groupe que la CE étudiée, la position dans la liste ainsi que le nombre cumulé de CE retrouvée sont reportés sur le graphique. Pour évaluer la classification de la CE, l'aire sous la courbe est calculée (AUC). Une AUC > 0.8 signifie une bonne classification de la CE.

Au regard des courbes individuelles des CE, l'estimation de la meilleure distance de classification est déterminée par le calcul de la moyenne des aires sous la courbe (AUC) individuelle de chaque CE (Figure 34). Si la distance euclidienne et la corrélation ont toutes les deux permis de classer les CE de façon uniforme (en termes de nombre de classes obtenues et de nombre de CEs par classe), il est en revanche intéressant de noter que la classification obtenue par la corrélation est plus précise et plus homogène que celle obtenue avec la distance euclidienne.

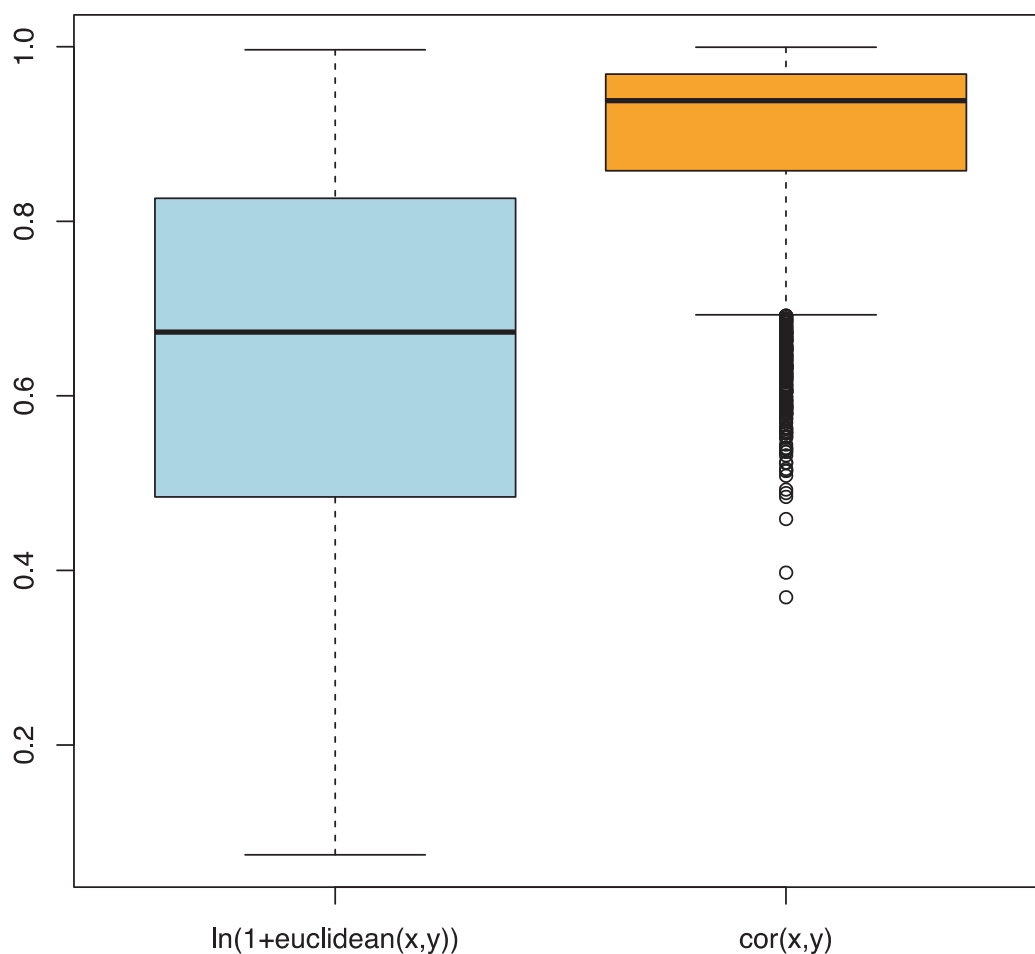


Figure 34. Estimation de la meilleure distance de classification

Boxplot de la moyenne des AUC des CE pour le log de la distance euclidienne (bleu) et la corrélation (orange). L'AUC individuelle de chaque CE est reportée pour le log de la distance euclidienne (bleu) et la corrélation (orange). La corrélation donne une classification plus précise et homogène que la distance euclidienne.

4.1.2.6. Sélection des groupes d'intérêts

La sélection de groupes de composés pertinents pour la perturbation endocrinienne et/ou la reprotoxicité a été faite via l'application de plusieurs critères : seuls ceux présentant une signature caractéristique comprise entre 10 et 120 gènes et regroupant entre 2 et 10 composés (et non CEs) associés significativement (sur-représentation) aux termes « Endocrine disorders », « Gonadal disorders » ou « Testicular diseases » ont été sélectionnés. Cette sélection permet de s'intéresser uniquement aux clusters les plus spécifiques, c'est-à-dire ceux qui vont potentiellement regrouper des molécules dont l'action est similaire. Elle a mis en évidence 3 groupes d'intérêts. Les courbes de classification des CEs de chacun de ces groupes sont disponibles en Annexe.

Le premier groupe sélectionné est constitué de 42 CEs, obtenues dans le foie et regroupées en utilisant la corrélation. Ce groupe est constitué de 8 molécules différentes correspondant à 42 CEs et présente une signature génique de 108 gènes. Ce groupe agglomère des médicaments à base d'hormones stéroïdiennes de synthèse ainsi que des hormones stéroïdiennes bien souvent utilisées dans les pilules contraceptives (Tableau XIII). Il est donc possible ici que les composés se soient regroupés sur la base d'une signature toxicogénomique caractéristique d'un impact sur la stéroïdogenèse ou de l'action sur les récepteurs stéroïdiens. Ceci peut être confirmé par la présence des gènes CYP3A18 (famille de cyP450) intervenant au sein de stéroïdogenèse (Nagata et al. 1996), CREM impliqué dans l'expression du gène StAR (Manna et al. 2002) et Paqr7 codant pour des récepteurs à la progestérone. La modification de la stéroïdogenèse étant souvent associée aux pathologies testiculaires, et dans un contexte d'identification de potentiels perturbateurs endocriniens, l'obtention de ce groupe est particulièrement intéressante.

Nom	MeSH ID	PE avéré	Type	Description
Ethinylestradiol	D004997	/	Médicament	Dérivé de synthèse de l'estradiol utilisé dans la majorité des pilules contraceptives
Methyltestosterone	D008777	/	Médicament	Forme synthétique de la testostérone utilisée pour traiter les hommes avec une déficience en testostérone
β-estradiol	D004958	/	Hormone	Dérivé de la testostérone nécessaire au maintien de la fertilité et des caractères sexuels secondaire chez la femme
Diethylstilbestrol	D004054	/	Médicament	Ostrogénique puissant utilisé dans les années 40 pour diminuer le nombre de fausse couches
Estriol	D004964	/	Hormone	Estrogène produit par le placenta
Fluconazole	D015725	/	Fongicide	Anti fongique utilisé pour traiter les infections
Mestranol	D008656	Danish EPA	Médicament	Estrogène de synthèse utilisé dans la pilule contraceptive
Norethindrone	D009640	/	Médicament	Progestatif de synthèse utilisé dans la pilule contraceptive

Tableau XIII. Molécules du premier groupe sélectionné.

Il est également à noter la présence d'un composé assez différent du reste du groupe : le fluconazole. Le fluconazole est un médicament antifongique de synthèse, apparenté à la famille des imidazoles. Son action antifongique s'exerce en inhibant une enzyme appartenant à la superfamille du cytochrome P-450 (14-déméthylase) entraînant la diminution de la synthèse de l'ergostérol, composant de la membrane cellulaire du champignon, indispensable à sa croissance. Récemment des études portées sur le fluconazole ont mis en avant l'utilité de cette molécule dans le traitement médicamenteux de la maladie de Cushing (Burns, Christie-David, and Gunton 2016). Cette maladie se développe lors d'une surproduction du cortisol par les glandes surrénales et peut être traitée soit par ablation de ces glandes soit par traitement chimique. Dans ce cadre, une étude a montré que le fluconazole inhibe la production de la 11- β hydroxylase, enzyme impliquée dans la stéroïdogénèse en catalysant la transformation du 11-deoxycortisol en cortisol. Cette étude a également mis en avant que le fluconazole exerce un effet sur le transporteur du cholestérol Star et l'enzyme de clivage de sa chaîne latérale la Cyp11A1. Au vu de ces effets, il est finalement cohérent que le fluconazole ait été regroupé au sein d'un groupe de molécules agissant sur la stéroïdogénèse. Par ailleurs, la courbe d'agglomération du fluconazole montre clairement que l'inclusion de cette molécule dans ce groupe n'est pas une erreur (Figure 35, courbe orange).

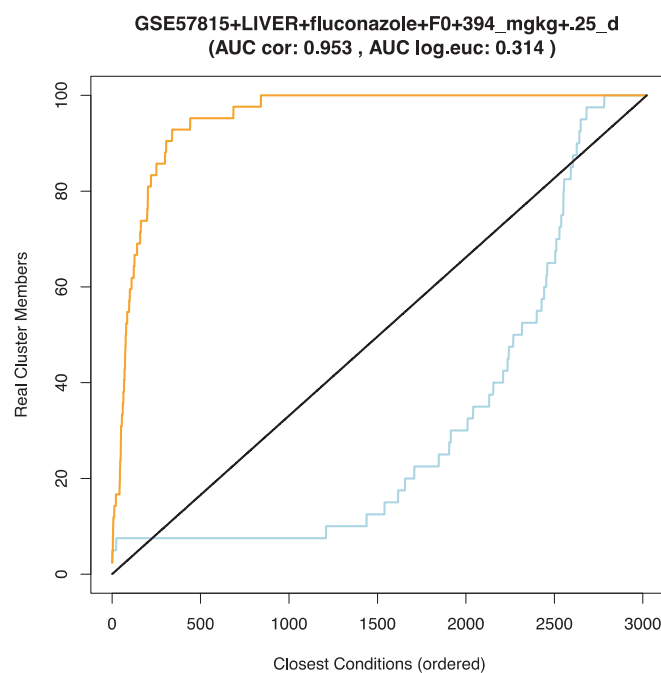


Figure 35. Classification du fluconazole à 394 mg/kg, 6 heures

Courbe d'agglomération de la CE impliquant le fluconazole à 394 mg/kg, 6 heures (selon la corrélation en orange et le log de la distance euclidienne en bleu). Ce groupe a été obtenu en utilisant la corrélation de Pearson. L'AUCcor > 0.8 de la courbe de corrélation indique que la CE n'a pas été classé dans ce groupe au hasard.

Notre deuxième groupe d'intérêt est constitué de 27 CEs obtenues dans le rein, et correspond à cinq composés (Tableau XIV).

Nom	MeSH ID	PE avéré	Type	Description
Clofibrate	D002994	/	Médicament	Hypolipémiant
Bezafibrate	D008777	/	Médicament	Hypolipémiant
Fenofibrate	D011345	/	Médicament	Hypolipémiant
Fénoprofène	D005279	/	Médicament	Anti-inflammatoire non stéroïdien
Gemfibrozil	D015248	/	Médicament	Hypolipémiant

Tableau XIV. Molécules présentes dans le second groupe

Ce groupe est représenté en majorité par des fibrates. Ces dernières années, les fibrates ont été sujets à de nombreuses investigations afin de déterminer leur potentiel de PE (Mimeault et al. 2005) et parmi les molécules sélectionnées, le bezafibrate a déjà été identifié comme potentiel PE (Hara et al. 2011; Velasco-Santamaría et al. 2011). En effet, les fibrates vont agir sur les récepteurs activés par les proliférateurs de peroxyosomes (PPARs) et plus spécifiquement les PPARs α . Ces derniers appartiennent à la superfamille des récepteurs nucléaires qui sont des facteurs de transcription activés par des ligands dont l'isoforme α joue un rôle dans la régulation du métabolisme des lipoprotéines et de la réponse inflammatoire (Desvergne and Wahli 1999). Si l'action des PPARs α n'est pas limpide, certaines études semblent mettre en avant une implication de ces récepteurs sur la stéroïdogenèse (Toda et al. 2003) et plus particulièrement sur les cellules de Leydig où ils suppriment le transport du cholestérol dans la mitochondrie (Gazouli et al. 2002).

À ces 4 fibrates est également associé un anti-inflammatoire non stéroïdien (AINS), le fénoprofène. Tout comme certains AINS, il s'agit d'un inhibiteur de la prostaglandine g/h synthétase ou cyclo-oxygénase, enzyme clé dans la biosynthèse des prostaglandines. Les prostaglandines sont des molécules de signalisation impliquées dans de nombreux processus métaboliques tels que la prolifération cellulaire, la différenciation, l'apoptose ou la réponse inflammatoire. Certaines études ont par ailleurs démontré que l'activation des PPARs α lors d'une inflammation pouvait être à l'origine d'une inhibition de la synthèse de molécules pro-inflammatoires telles que les prostaglandines (Poynter and Daynes 1998). Cela pourrait expliquer le fait que le fénoprofène puisse être associé à la perturbation endocrinienne et soit regroupé au sein d'un groupe fibrates. Ici encore, après vérification de la courbe d'agglomération, le regroupement du fénoprofène avec les autres composés du groupe apparaît clairement ne pas être dû au hasard (Figure 36, courbe orange).

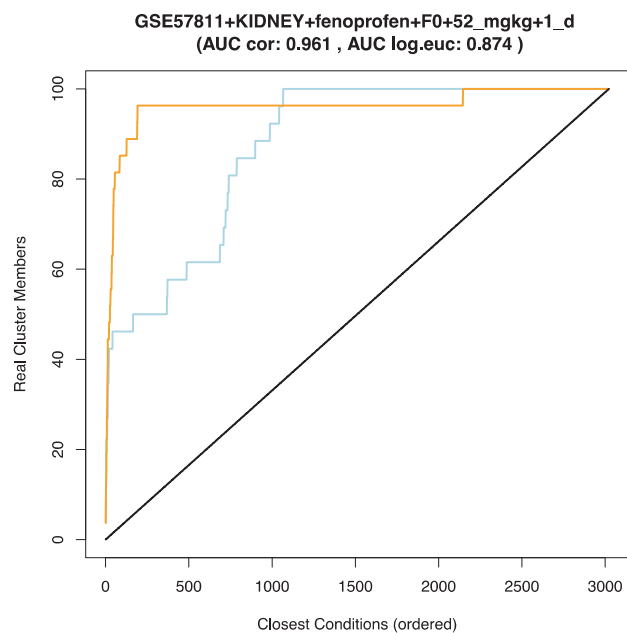


Figure 36. Courbe de classification du fénoprofène à 52 mg/kg, 1 jour

Courbe d'agglomération de la CE impliquant le fénoprofène à 52 mg/kg, 1 jour (selon la corrélation en orange et le log de la distance euclidienne en bleu). Ce groupe a été obtenu en utilisant la corrélation de Pearson. L'AUCcor > 0.8 de la courbe de corrélation indique que la CE n'a pas été classé dans ce groupe au hasard.

Enfin le troisième groupe, plus hétérogène au niveau de ses composés, regroupe neuf molécules pour 21 CEs obtenues dans le foie avec la distance euclidienne. Ce groupe rassemble plusieurs composés ne présentant pas un mécanisme d'action partagé. Si certaines molécules ont tendance à agir de la même manière, comme la phénacétine et l'acétaminophène par exemple, les autres substances ont a priori des mécanismes propres. (Tableau XV).

Nom	MeSH ID	PE avéré	Type	Description
Acétaminophène	D000082	/	Médicament	Antalgique le plus utilisé dans le monde
Bromobenzène	D001969	/	Composé organique	Hydrocarbure bromé utilisé comme additif dans l'essence
Butylated hydroxyanisole	D002083	TEDX	Additif alimentaire	Utilisé comme anti oxydant dans l'industrie alimentaire
Chlormezanone	D002720	/	Médicament	Anxiolytique
Dantrolène	D003620	/	Médicament	Myorelaxant
Ethionamide	D005000	/	Médicament	Antibiotique utilisé contre la tuberculose
Flutamide	D005485	/	Médicament	Anti-androgène non stéroïdien utilisé pour le cancer de la prostate
Methimazole	D008713	/	Médicament	Antithyroidien
Phénacétine	D010615	/	Médicament	Analgésique

Tableau XV. Molécules présentes dans le troisième groupe

Malgré son hétérogénéité apparente, ce groupe est particulièrement intéressant, car sur les 9 molécules dont il est composé, le butylated hydroxyanisole est un PE avéré tandis que quatre autres sont des PE suspectés et/ou en cours d'étude dans notre laboratoire :

Acétaminophène : plusieurs études épidémiologiques ont montré l'impact de l'acétaminophène chez la femme enceinte sur la prévalence d'une cryptorchidie chez les nouveau-nés (Snijder et al. 2012; Jensen et al. 2010; S?verine Mazaud-Guittot et al. 2013) ainsi que son effet anti-androgène sur le testicule (Kristensen et al. 2012).

Butylated hydroxyanisole : l'effet PE de cet antioxydant n'est pas assez documenté pour affirmer avec exactitude s'il possède une action sur le SE ou non (Pop, Kiss, and Loghin 2013). Cependant, de forts soupçons pèsent sur cette molécule quant à son effet (anti) oestrogénique et (anti) androgénique (Kang et al. 2005; Jobling et al. 1995).

Ethionamide : plusieurs études de cas ont mis en évidence la possible implication de cette molécule dans le développement d'hypothyroïdies sûrement dues à son action sur la *Thyroid stimulating hormone* (Drucker et al. 1984; Mallela et al. 2015).

Flutamide : l'action du flutamide sur les récepteurs des androgènes présents au niveau de la prostate en fait un potentiel PE (Sogani, Vagaiwala, and Whitmore 1984). Ceci est d'autant plus problématique que de faibles doses de cette molécule peuvent être retrouvées dans les eaux de surfaces (Shi et al. 2016). Enfin, certaines études imputent au flutamide l'apparition de pathologies génitales mâles comme l'hypospadias (Sinclair et al. 2017).

Methimazole : ce médicament utilisé en cas d'hyperthyroïdie dont l'action goitrigène (perturbation des hormones thyroïdiennes) est bien connue. Il agit directement sur la synthèse d'hormones thyroïdiennes et plus particulièrement sur la *thyroid peroxidase*, enzyme majeure dans la conversion de la tyrosine en hormones thyroïdiennes T3 ou T4 (Song et al. 2012).

Cependant il faut noter que parmi les 21 CE composants ce groupe, seule une seule présente une AUC au-dessus de 0.8. La moyenne du groupe se trouvant à 0.712 et la médiane à 0.714. Cela peut s'expliquer par le fait que le groupe a été aggloméré en utilisant le log de la distance euclidienne, métrique présentant les groupes les plus hétérogènes.

Malgré des AUC inférieures à notre seuil, on peut dire que les CE n'ont pas été regroupées au hasard. Qui plus est, cette classe comporte de nombreux perturbateurs endocriniens suspectés, c'est pourquoi j'ai décidé de garder ce groupe dans ma sélection.

4.1.2.7. Validations expérimentales

Nous travaillons actuellement à l'évaluation expérimentale de 22 substances extraites de la sélection de groupes dont le bromobenzène, le chlomezanone, le dantrolène et le phénacétine molécules extraites du troisième groupe de sélection et associés avec des PE.

Cette validation s'inscrit dans le cadre du projet MASSIVE ATTACK, lui-même dans la continuité directe de ChemPSy. Il ambitionne d'étudier l'apport des signatures métabolomiques et transcriptomiques dans l'évaluation des risques sanitaires cumulés. Il s'agit d'évaluer les performances de classification et de priorisation de ChemPSy en y intégrant de nouvelles données transcriptomiques et métabolomiques non ciblées d'une centaine de molécules (dont les 22 sélectionnées) par l'utilisation de cultures cellulaires, en particulier la lignée NCI-H295R.

La première étape de validation consiste à déterminer la dose minimale non cytotoxique des molécules par criblage *in vitro* uniformisé sur un plateau automatisé de cultures et de tests cellulaires (plateforme ImPACcell, Rennes). Chaque molécule sera testée à plusieurs concentrations et une analyse métabolomique sera réalisée sur la dose minimale non cytotoxique afin de déterminer la présence d'un effet sur la testostérone. L'impact sur la sécrétion de testostérone par la lignée NCI-H295R sera le critère

principal retenu comme indicateur de perturbation endocrinienne. Les molécules n'ayant aucun effet PE sont retirées du set de substances à tester.

La seconde étape consistera au même criblage *in vitro* des molécules restantes a minima à 6 doses et en quadruplicat selon le protocole de l'OCDE (Ligne Directrice n°456). Puis, la signature métabolomique et (notamment stéroïdomique) sera ensuite déterminée et analysée dans les milieux de culture (Laboratoire d'étude et de recherche en environnement et santé – LERES, Rennes) afin de modéliser les relations doses-réponse pour chaque molécule. De ces courbes, la concentration effective et notamment la EC50, concentration de la substance entraînant une modification de réponse de 50 % par rapport au témoin dans une expérimentation donnée, seront estimées.

Les molécules présentant un effet de perturbation endocrinienne *in vitro* seront à nouveau testées en triplicat sur la lignée NCI-H295R, aux CE50s retenues. Les ARNs seront extraits pour l'analyse toxicogénomique et les milieux de culture prélevés pour l'analyse métabolomique finale. Les ARNs extraits permettront de définir la signature toxicogénomique de chaque substance en utilisant la technologie du 3' *Digital Gene Expression Sequencing* (3' DGE-seq).

4.1.2.8. Prédiction des effets de nouveaux composés

L'objectif ici est de tester notre classification et de déterminer une méthode pour prédire à quel groupe appartient un nouveau composé sur la base de sa signature toxicogénomique. À défaut d'utiliser une signature extérieure au jeu de données, j'ai utilisé les CE déjà présentes et tenté de les reclasser selon différentes méthodes. Pour chacune des quatre matrices, la sensibilité, la spécificité, le taux d'erreur ainsi que le taux de réussite (1-taux d'erreur) ont été obtenues (centroïde, médoïde, et par CE la plus proche). Pour les méthodes de SVM et de PLS, uniquement la sensibilité et la spécificité ont été obtenues.

Le Tableau XVI résume la spécificité et la sensibilité moyenne pour chaque technique, matrice et distance. Cette moyenne est obtenue par le calcul de la sensibilité et de la spécificité moyenne de chaque groupe et pour chaque randomisation. Si les valeurs sont proches les unes des autres, on note tout de même qu'en moyenne les méthodes basées sur le centroïde et le médoïde présentent de meilleurs résultats en terme de spécificité. Pour ce qui est de la sensibilité, les méthodes de SVM, PLS et de centroïde présentent les meilleurs résultats. On note également qu'il est plus facile de détecter les vrais positifs et les vrais négatifs sur les matrices log2-(fold-change) en utilisant la corrélation. De même, la méthode de reclassification en fonction des CE les plus proches présente de moins bons résultats avec la distance euclidienne, par rapport aux techniques de centroïde et médoïde, et aux techniques utilisant

la corrélation. Ceci s'explique par le fait que lors de la classification, une CE peut avoir été associée à un groupe plus par défaut que par similarité. Cette CE n'est pas représentative du groupe et peut être, lors d'une représentation spatiale, très éloignée du centroïde de sa classe.

A

Spécificité	Distance Euclidienne				Corrélation de Pearson				Moyenne par méthode
	Matrice discrétisée	Matrice lof2fc	Matrice PCA discrétisée	Matrice PCA log2fc	Matrice discrétisée	Matrice lof2fc	Matrice PCA discrétisée	Matrice PCA log2fc	
SVM	0.858	0.846	0.859	0.846	0.832	0.832	0.832	0.832	0.842125
PLS	0.801	0.807	0.801	0.807	0.799	0.814	0.799	0.814	0.80525
Centroïde	0.991	0.991	0.991	0.991	0.998	0.997	0.997	0.997	0.994125
Médoïde	0.991	0.991	0.991	0.991	0.996	0.996	0.998	0.997	0.993875
CEs la plus proche	0.990	0.990	0.990	0.990	0.998	0.998	0.998	0.998	0.994
Moyenne par matrice	0.9262	0.925	0.9264	0.925	0.9246	0.9274	0.9248	0.9276	

B

Sensibilité	Distance Euclidienne				Corrélation de Pearson				Moyenne par méthode
	Matrice discrétisée	Matrice lof2fc	Matrice PCA discrétisée	Matrice PCA log2fc	Matrice discrétisée	Matrice lof2fc	Matrice PCA discrétisée	Matrice PCA log2fc	
SVM	0.940	0.932	0.935	0.931	0.939	0.946	0.939	0.944	0.93825
PLS	0.960	0.950	0.960	0.950	0.942	0.938	0.942	0.937	0.947375
Centroïde	0.878	0.868	0.871	0.877	0.930	0.927	0.926	0.922	0.899875
Médoïde	0.879	0.878	0.884	0.879	0.908	0.910	0.911	0.910	0.894875
CEs la plus proche	0.839	0.839	0.837	0.836	0.903	0.903	0.909	0.902	0.871
Moyenne par matrice	0.8992	0.8934	0.8974	0.8946	0.9244	0.9248	0.9254	0.923	

Tableau XVI. Spécificité et sensibilité moyenne pour chaque technique de classification

Pour chaque matrice et chaque distance, la spécificité (table A) et la sensibilité (table B) des groupes sont déterminées en fonction de la technique utilisée (SVM, PLS, centroïde, médoïde et CEs la plus proche). La moyenne des sensibilités et spécificités individuelle des groupes est ensuite calculée et reportée dans le tableau.

Les courbes d'erreurs (Figure 37) et de réussite (Figure 38) permettent de se rendre compte de la robustesse des groupes ainsi que des méthodes de classification. En classifiant de nouveau les CEs, le groupe associé à la prédiction est bien celui d'origine de la CE dans 73% des cas pour la corrélation et dans 56% pour la distance euclidienne, toutes matrices et méthodes confondues. Le groupe d'origine de la CE, si ce n'est pas le groupe le plus proche, se trouve dans les 5 premiers groupes, dans 92% des cas pour la corrélation et dans 76% des cas pour la distance euclidienne, toutes matrices et méthodes confondues ici aussi.

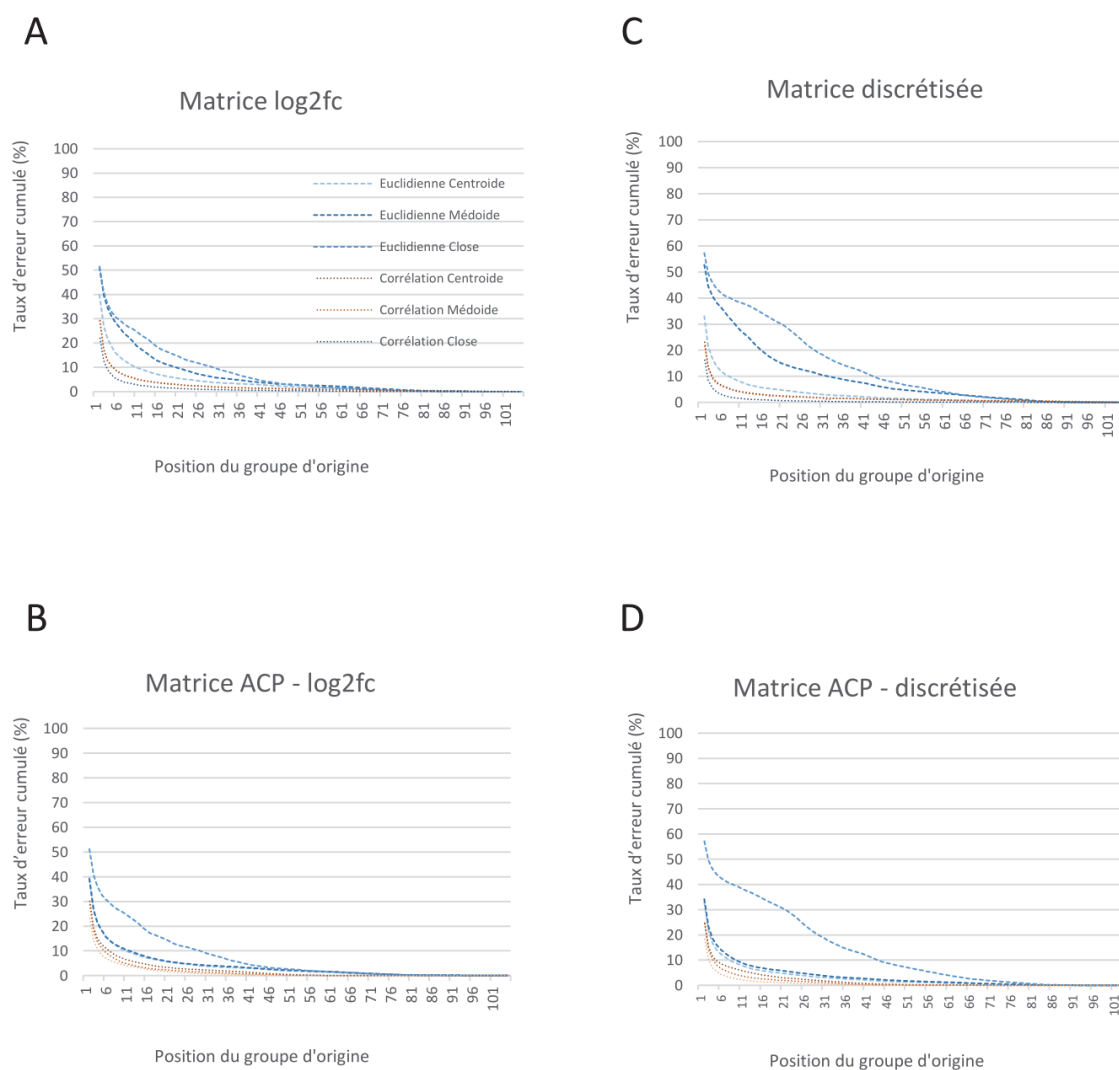


Figure 37. Courbe d'erreur cumulée de la reclassification

Pour chaque distance (euclidienne : bleue, corrélation : orange) le taux d'erreur de reclassification a été calculé en fonction de la position du groupe d'origine en utilisant soit la matrice log2fc (A), la matrice ACP-log2fc (B), la matrice discrétisée (C), la matrice ACP-discrétisée (D)

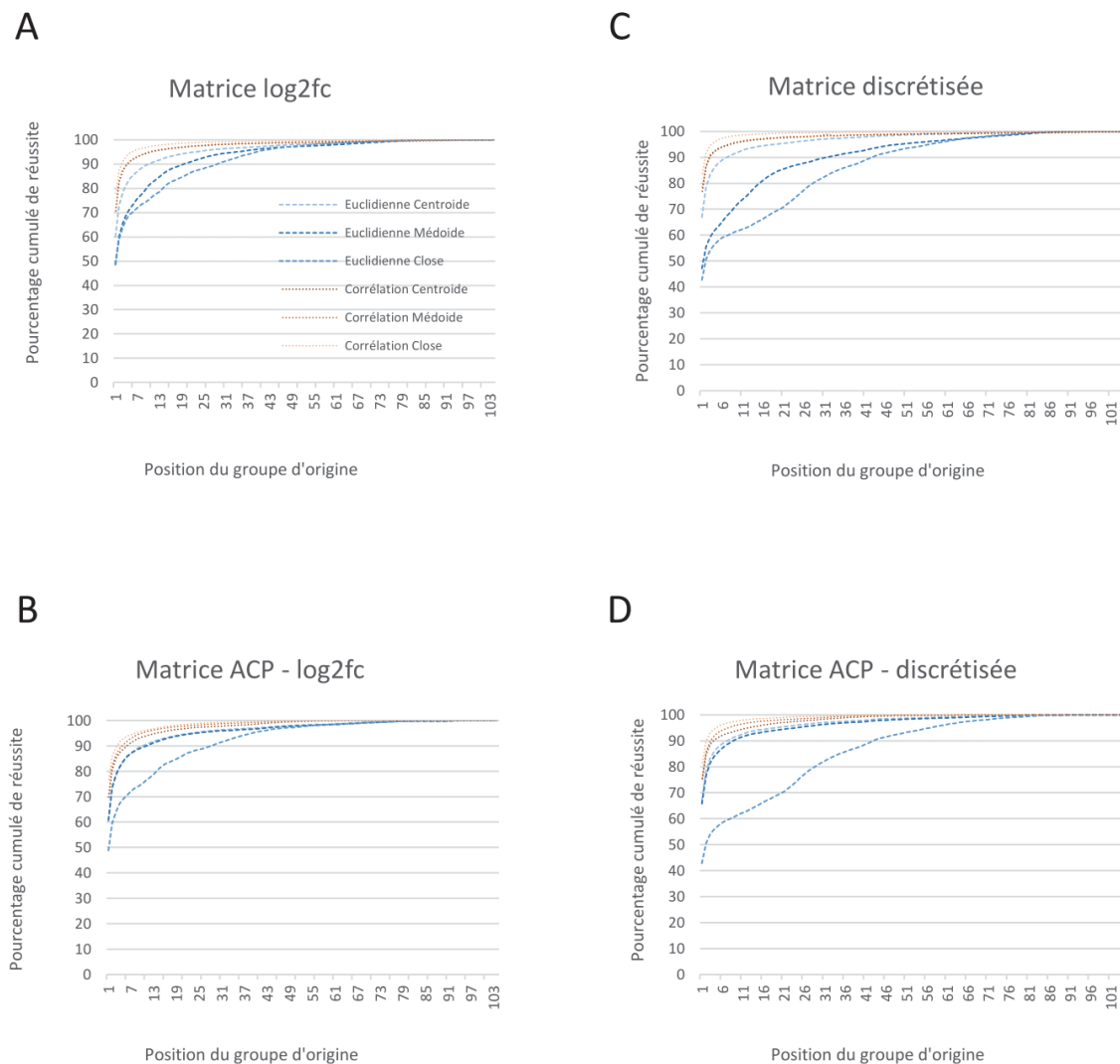


Figure 38. Courbe de pourcentage de réussite cumulé de la reclassification

Pour chaque distance (euclidienne : bleue, corrélation : orange) et pour chaque méthode, le taux de réussite de reclassification est représenté en fonction de la position du groupe d'origine des CEs en utilisant soit la matrice log2fc (A), la matrice ACP-log2fc (B), la matrice discrétisée (C), la matrice ACP-discrétisée (D).

À partir de la classification des données obtenues et en fonction des matrices, il est possible par l'intermédiaire de ces différentes méthodes, en utilisant de préférence la corrélation sur les matrices log2-(fold change), de prédire le groupe le plus proche de la signature toxicogénomique d'un nouveau composé et ainsi lui inférer les phénotypes sur-représentés dans ce groupe.

Il faut toutefois nuancer les résultats obtenus par la SVM et la PLS. En effet ces techniques nécessitent un jeu d'apprentissage pour définir les bornes du groupe. Aussi, récupérer 20% des CEs de chacune de ces classes peut devenir problématique, notamment quand le groupe est composé de peu de CEs. Si cela n'empêche, pas dans la majorité des cas, la SVM de classer correctement la CE de test, l'intervalle de confiance à 95% de cette classification peut être compris, en fonction du groupe, entre 2,5% et 27%. Pour éviter ce type de problèmes, il est nécessaire d'utiliser un jeu de données plus important.

4.1.3. Conclusion et perspectives du projet ChemPSy

Le processus d'évaluation des risques chimiques est aujourd'hui un challenge pour les autorités compétentes et implique un fort investissement de temps et d'argent pour pointer du doigt les effets néfastes des substances chimiques. Dans ce cadre, l'utilisation de la toxicologie computationnelle présente de nombreux avantages pour orienter ces évaluations. Elle est actuellement utilisée pour le développement de signatures prédictives (Martin et al. 2011). Dans cette optique, le projet ChemPSy a permis de regrouper des composés chimiques sur la base de leurs signatures toxicogénomiques. L'utilisation des méthodes développées au sein de ce projet peut être le point de départ d'une nouvelle approche pour l'aide à la décision afin d'orienter l'évaluation des risques chimiques. Il a de ce point de vue donné pleine satisfaction et s'est montré riche d'enseignements. D'une part, celui-ci nous a permis de réaliser que les données disponibles dans les banques de toxicogénomique, telles que la CTD, ne permettaient pas de développer des approches prédictives efficaces (sur la base des méthodologies testées). D'autre part il a mis en évidence l'importance et les limitations de la fouille de données et de leur standardisation. La constitution du jeu de données quantitatives représente la première limitation. En effet pour sa mise en place plusieurs dépôts de données ont dû être consultés, tous publics, mais pas forcément tous connus. Si les données issues de l'analyse DrugMatrix sont facilement accessibles, celles d'Open TG-GATEs le sont beaucoup moins. Il a fallu les télécharger depuis le site du consortium TG-GATEs, site de base en japonais. Il est alors facile de comprendre le problème rencontré par les biologistes face à ce stockage épars des informations. La deuxième limitation soulevée et sans doute la plus importante est celle de la standardisation des données. Après extraction des informations d'intérêt, il est en effet nécessaire de supprimer les informations redondantes et de compléter celles manquantes (Audouze and Taboureau 2015). Cette tâche difficilement automatisable est d'autant plus difficile à mettre en place que la plupart du temps ces informations ne sont pas uniformisées. En effet, le vocabulaire utilisé pour décrire les données diffère d'une expérience à une autre. En toxicologie, l'exemple le plus flagrant est celui du nom des composés chimiques. Selon les publications, le nom usuel de la molécule pourra être utilisé, mais il n'est pas rare de retrouver plutôt son numéro CAS (*Chemical Abstracts Service*) ou encore son synonyme. Par exemple l'estradiol peut être présenté sous le beta-estradiol, β -estradiol, 17beta-estradiol, 17 β -estradiol, Œstradiol ou encore par son CAS 50-28-2. Face à la multiplication de ces problèmes, de nombreuses solutions sont mises en place que ce soit pour la standardisation des composés chimiques (Eltyeb and Salim 2014), des gènes (Yates et al. 2017) mais également de l'activité biologique des substances (Mathias et al. 2013).

Si les premiers résultats sont encourageants, il reste de nombreuses améliorations à apporter à la méthodologie développée. La première consiste à uniformiser les signatures. En effet, lors de la classification, nous avons pu observer une tendance à regrouper les CEs non pas en fonction de la signature du composé, mais en fonction de l'organe d'origine de la signature. Ceci explique pourquoi dans les classes que j'ai obtenues les CEs d'un seul organe est toujours retrouvé. L'uniformisation des signatures (par l'analyse d'un seul organe ou type cellulaire) permettrait de gommer l'effet "organe" pour se focaliser uniquement sur les signatures des composés. Néanmoins, en utilisant des filtres sélectifs comme ceux appliqués dans notre sélection, il est possible d'observer un regroupement par mode action et/ou par similarité des substances. Par ailleurs, et en dépit du poids des organes dans la classification, il s'est avéré possible d'associer à des groupes de CEs obtenues dans un organe donné, le foie par exemple, des effets toxicologiques ou pathologies en lien avec un tout autre organe, comme ici le testicule. Ceci met clairement en évidence l'existence de signatures caractéristiques des composés chimiques qui, indépendamment de l'organe analysé, permettent la prédiction de leurs effets toxicologiques. Enfin, les méthodes de classifications cherchent à regrouper toutes les CEs, quitte à agglomérer des CEs qui ne feront qu'augmenter la variabilité intragroupe. Afin d'obtenir des groupes plus homogènes il faudrait supprimer dans chaque groupe les CEs présentant une AUC inférieure à 0,8. Cette élimination concernerait environ 30% des CEs agglomérés sur la base de la distance euclidienne et 10% en utilisant la corrélation.

La seconde amélioration nécessaire concerne évidemment la taille de notre jeu de données : Plus celui-ci sera grand et diversifié, en termes de composés chimiques, mais surtout en termes de modes d'action de ceux-ci, plus la classification sera précise et donc meilleures seront les prédictions d'effets toxicologiques. Aujourd'hui ChemPSy comprend deux des trois plus gros jeux de données toxicogénomiques sur puces à ADN, à savoir Open TG-GATEs et DrugMatrix. Notre base de travail peut notamment être agrandie par l'intégration de *The Connectivity Map* (CMAP) (Lamb et al. 2006), un programme initié par le *Broad Institut* et qui vise à analyser la réponse toxicogénomique de 164 composés dans différentes lignées cellulaires humaines. Dans cette perspective et afin de répondre aux limitations évoquées précédemment, le projet MASSIVE ATTACK a été mis en place. Celui-ci vise à analyser la réponse toxicogénomique de plus d'une centaine de composés qui viendront donc consolider notre jeu de signatures. Par ailleurs, l'intégration de données de métabolomique viendra ajouter permettra d'affiner le clustering des CEs en fonctions des résultats de métabolomique.

Enfin, ChemPSy repose sur les interactions gènes-pathologies décrites dans la CTD pour prédire les effets toxicologiques potentiels d'un groupe de composé. La CTD n'étant pas exhaustive, l'intégration des associations composés-pathologies avec une autre source d'information devra être

envisagée. Parmi les solutions possibles, l'utilisation du protocole expérimental utilisé dans la conception de la base de données ChemProt (Kringelum et al. 2016) est favorisée. Brièvement, cette banque de données repose sur la mise en relation de pathologies avec des composés chimiques, par extrapolation à partir des annotations des protéines. Celles-ci sont obtenues par l'agrégation d'un grand volume de données extraites de banques d'interaction protéines-pathologies tel que OMIM (Amberger et al. 2015), GeneCards (Safran et al. 2002), *Human Protein Atlas* (Uhlen et al. 2015). En utilisant la description des complexes protéines-pathologies il sera alors possible d'obtenir les associations substances chimiques-pathologies. L'intégration de la base ChemProt pour améliorer les associations entre composés chimiques et pathologies serait de fait une piste intéressante dans le but d'améliorer les effets toxicologiques prédits par ChemPSy.

Les problèmes rencontrés lors de la constitution du jeu de données, à savoir l'extraction des signatures toxicogénomiques, mais également la description précise et homogène des différentes CE, nous ont fait réaliser l'absence cruciale d'espaces de dépôts spécialisés dans le domaine de la toxicologie, particulier en toxicogénomique. De ce constat a découlé TOXsIgN, le second projet de ma thèse.

4.2. TOXsIgN : un espace de dépôt pour les signatures toxicogénomique

Face au nombre croissant de composés chimiques dans notre environnement, les expérimentations visant à déterminer les effets délétères de ces substances se sont considérablement intensifiées. Or, de nombreuses inquiétudes ont récemment été soulevées à propos du manque de reproductibilité des données dans les sciences de la vie, et plus particulièrement dans le domaine de la toxicologie (“Must Try Harder” 2012; Miller 2014; Poland et al. 2014). Le *National Institutes of Health* (NIH) et d’autres organismes de financement ou de protection partagent également cette préoccupation quant à la reproductibilité des données en sciences de l’environnement (Collins and Tabak 2014). Une des solutions envisagées consiste en la mise à disposition et en une plus grande transparence des données générées. Si la toxicogénomique a significativement contribué à améliorer les connaissances sur les mécanismes de toxicité des composés chimiques, le stockage et l’accessibilité des données générées restent un problème.

Dans ce contexte, de nombreuses banques de données ont vu le jour. Parmi eux, CEBS (Lea et al. 2017), ToxBank (Kohonen et al. 2013) ou encore diXa (Hendrickx et al. 2015) offrent des espaces dédiés pour déposer, annoter et structurer les données brutes issues d’analyses toxicogénomiques. De son côté, ToxCast (US EPA) a été développé dans le but de prédire la toxicité d’un composé chimique à partir de son profil de bioactivité (Richard et al. 2016). De même, LINCS (NIH) correspond à un consortium visant à améliorer les connaissances en biologie en cataloguant les changements transcriptionnels des cellules lorsque celles-ci sont exposées à des agents perturbateurs (Cheng and Li 2016). Ces deux derniers programmes mettent à disposition un ensemble d’outils permettant la consultation de leurs données, mais également la prédiction de toxicité ou encore l’extraction de signatures toxicogénomique. Cependant, aucune de ces ressources web ne permet à la communauté scientifique de déposer le résultat d’analyses sous forme de signatures toxicogénomiques – c’est-à-dire l’ensemble des effets génomiques sur un individu ou sa descendance après exposition à des facteurs environnementaux (simples ou en mélanges) : chimiques (pesticides, plastifiants...), physiques (radiations, température...) ou biologiques (agents pathogènes, parasites).

Si aujourd’hui, les éditeurs imposent aux auteurs la soumission de leurs données brutes au sein d’espaces de dépôts tels que GEO ou ArrayExpress (Barrett et al. 2013; Kolesnikov et al. 2015), les données analysées sont souvent présentées sous un format inexploitable (PDF) dans les « Supplementary information » des papiers. La majeure partie du temps, il est donc impossible pour les scientifiques d’avoir accès à ces résultats et de pouvoir comparer leurs résultats avec ceux publiés. De plus,

l'hétérogénéité et la faible description de ces données « brutes » rendent le processus de comparaison encore plus difficile et imposent que leur traitement soit réalisé par des scientifiques ayant de solides bases bio-informatiques et biostatistiques. Face à ce constat, nous avons entrepris de développer un espace de dépôt dédié à la soumission de signatures toxicogénomiques : TOXsIgN (*TOXicogenomics sIgNatures*). TOXsIgN offre la possibilité aux toxicologues et toxicogénomistes d'entreposer le résultat de leurs analyses sous la forme de signatures. Ces dernières se composent de deux listes de gènes, une pour les gènes surexprimés et une autre pour les gènes sous-exprimés en réponse à une exposition à un facteur environnemental. Chaque information associée à ces signatures est décrite par l'utilisation de vocabulaires contrôlés. Enfin, TOXsIgN intègre également un espace de travail hébergeant des outils bio-informatiques afin d'analyser et d'interpréter les signatures toxicogénomiques déposées.

4.2.1. Manuscrit en préparation

Databases and ontologies

TOXsigN: a cross-species repository for toxicogenomics signatures

Thomas A. Darde^{1,2}, Pierre Gaudriault¹, Rémi Beranger¹, Clément Lancien¹, Annaëlle Caillarec-Joly¹, Olivier Sallou², Nathalie Bonvallot^{1,3}, Nathalie Costet¹, Cécile Chevrier¹, Séverine Mazaud-Guittot¹, Bernard Jegou^{1,3}, Olivier Collin², Emmanuelle Becker¹, Antoine D. Rolland¹ and Frédéric Chalmel^{1,*}

¹ Inserm U1085-IRSET, Université de Rennes 1, 9 Avenue du Professeur Léon-Bernard, F-35000 Rennes, France ; ² Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA/INRIA) - GenOuest platform, Université de Rennes 1; F-35042 Rennes, France ; ³ EHESP – School of Public Health, 9 Avenue du Professeur Léon-Bernard, F-35000 Rennes, France.

* To whom correspondence should be addressed: Tel: +33 (0)2 23 23 58 02; Email: frederic.chalmel@inserm.fr

Email addresses:

TD: thomas.darde@inria.fr

PG: pierre.gaudriault@gmail.com

RB: remi.beranger@univ-rennes1.fr

CL: clement.lancien@gmail.com

OS: olivier.sallou@irisa.fr

ACJ: annaëlle.caillarec-joly@etudiant.univ-rennes1.fr

NC: nathalie.costet@univ-rennes1.fr

NB: nathalie.bonvallot@ehesp.fr

CC: cecile.chevrier@univ-rennes1.fr

SMG: severine.mazaud@univ-rennes1.fr

BJ: bernard.jegou@inserm.fr

OC: olivier.collin@irisa.fr

EB: emmanuelle.becker@univ-rennes1.fr

ADR: antoine.rolland@univ-rennes1.fr

FC: frederic.chalmel@inserm.fr.

Abstract

Motivation: While considerable worries are being raised about the lack of reproducibility in the field of toxicology, several databases have already paved the way for improving storage, exchange and analysis of raw toxicogenomics data. However, none of them provides access to processed and filtered data as originally reported in scientific publications. Given the increasing demand for accessing such information, we developed TOXsIgN, a repository for TOXicogenomics sIgNatures.

Results: The TOXsIgN repository provides a flexible environment to facilitate online submission, storage and retrieval of toxicogenomics signatures by the scientific community. It currently hosts 754 projects that describe more than 450 distinct chemicals and their 8491 associated signatures. Additionally, TOXsIgN provides users with a working environment containing a powerful search engine as well as bio-informatics/-statistics modules enabling enrichment analyses or signature comparisons.

Availability and Implementation: The TOXsIgN repository is freely accessible at <http://toxsign.genouest.org>. Website implemented in Python, JavaScript and MongoDB, with all major browsers supported.

Contact: frederic.chalmel@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The modern way of life results in humans to be exposed to numerous environmental components. As of July 2017, over 130 million chemicals were listed in the Chemical Abstract Service (CAS) among which 100,000 are manufactured across the world according to the European Inventory of Existing Commercial Chemical Substances (EINECS). The potential toxicity of the overwhelming majority of these compounds remain unknown (1). In 2007, the growing concerns about their potential adverse effects on human health and the environment have led the European Commission to create the Regulation on Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) (2). Such regulation promotes innovative scientific programs aiming at: *i*) screening novel potential toxicants to which humans are exposed, in particular the so-called CMR (carcinogenic, mutagenic, reprotoxic) substances; *ii*) investigating the molecular mechanisms underlying their actions; and, finally, *iii*) developing predictive methodologies to assess chemical hazard which should ultimately reduce the number of experimental tests on model organisms (3). In this context, several landmark projects, such as Open TG-GATEs (4), DrugMatrix (5), CMap (6) and a myriad of other high throughput studies, have been undertaken based on the assumption that toxicogenomics data, i.e. gene expression profiles in response to chemicals, are expected to improve predicting and understanding their mechanisms of toxicity (7). This concept thus supplements the traditional ligand- and structure-based predictive approaches to assess the safety of compounds by postulating that transcript profiling can effectively discriminate classes of compounds showing similar adverse effects (8, 9).

Recently, considerable worries have been raised regarding the lack of reproducibility of biomedical research (10), and more particularly in the field of toxicology (11–13). Funding agencies, such as the National Institutes of Health (NIH), share this concern and discuss ways to enhance reproducibility in environmental sciences (14). Among all practical points to consider when planning and reporting toxicology studies, one of them, raised by Collins and colleagues, is to provide a greater transparency of the data, including negative findings or contradictory results (13). Some agencies, such as the European Research Infrastructure Consortium (ERIC), try to establish a model service for system biology data management. The objectives are to make biological data FAIR: Findable, Accessible, Interoperable and Reusable (15). This is especially the case for toxicology. Several resources, such as CTD (16) (<http://ctdbase.org>), diXa (17) (<http://www.dixa-fp7.eu>), ToxDB (18) (<http://toxdb.molgen.mpg.de>), CEBS (19) (<https://www.niehs.nih.gov/research/resources/databases/cebs/index.cfm>) and Drug2Gene (20) (<http://www.drug2gene.com>) have paved the way for improving toxicological data storage, exchange and analysis (21). The Toxicity Forecaster (ToxCast) developed by the US Environmental Protection Agency is another good example of innovative tools aiming at generating and sharing standardized data and predictive models for thousands of toxicants, including endocrine disruptors (22).

It is now well-established that scientists must make their data publically available, especially when it comes to omics experiments, at the time they publish their results. Nowadays, however, investigators are only supposed to submit raw data in public databases such as the Gene Expression Omnibus (23) and ArrayExpress (24). Nevertheless, to the best of our knowledge, none of the resources mentioned

above allows scientists to actively submit toxicogenomics signatures, i.e. the sets of genes showing altered expression in individuals or their descendants after exposure to single or combined environmental factors. While those signatures constitute the heart of toxicogenomics studies, they usually only appear as supplementary tables which further complicate their direct reusability and comparison. Given the demand of scientists for easily accessing such information (i.e. without a time-consuming downloading and re-processing raw data), we developed TOXsIgN (for TOXicogenomics sIgNatures), a user-friendly resource that supports online submission, storage and retrieval of toxicogenomics signatures. One of the unique features of TOXsIgN relies on its ability to archive heterogeneous data as it allows users to upload lists of over-/under-expressed genes from different kinds of omics experiments (e.g. transcriptomic, proteomic, epigenomic) while making them compatible with cross-species and cross-technology comparisons. Furthermore, TOXsIgN also intends to serve as a warehouse for toxicogenomics and predictive toxicology tools leaning on the overall set of signatures deposited by the community.

2 Method

2.1 Data storage, management and retrieval

The database underlying TOXsIgN is based on MongoDB (<https://www.mongodb.com>), a free and open-source cross-platform document-oriented database program. This NoSQL database technology provides many interesting features including flexible storage of massive and rapidly changing types of data, data replication and JavaScript compatibility.

The TOXsIgN search engine is based on the implementation of an ElasticSearch server (<https://www.elastic.co>). Briefly, ElasticSearch is a NoSQL database manager with a powerful search engine primarily used to index textual data. It thus allows TOXsIgN to index all four layers of a TOXsIgN project information (project, study, assay, and signature) and investigators to query the database by using several entry types such as chemicals, genes, doses, species, cell lines and tissues.

2.2 Web interface

The web interface of TOXsIgN was built using two open web frameworks, Pyramid (<https://trypyramid.com/>) and AngularJS (<https://angularjs.org/>). The former embeds many features such as a REST API, a JSON renderer, a SQLAlchemy Object-relational mapper (ORM), a Deform library for forms generation, and a compatibility with SMTP server. The later corresponds to an open JavaScript framework which is an extension of traditional HTML vocabulary, allowing implementation of readable and quick to develop web environments.

To ensure website traffic as well as data security, scalability and deployment, each component of TOXsIgN (website-server, MongoDB database, Elasticsearch server) are hosted on an individual virtual machine using Docker (<https://www.docker.com>) as a container system. Briefly, Docker is an open-source virtualization system allowing easy deployment of container images which correspond to

lightweight, stand-alone, executable packages embedding everything needed to run a piece of software, such as code, runtime, system tools, libraries and settings.

2.3 Workspace

The workspace leans on the database and includes a job submission system to execute local scripts written in Python, R and Tcl/Tk, and any kind of stand-alone programs. The corresponding web interface is based on the Pyramid framework as well as Python scripts and allows users to run modules and to access the corresponding results. When a module is executed by the investigator, a new job is created by using a python wrapper. The status of each job is available in a table accessible through the “running jobs” web page in which queued, running and complete jobs are colored in grey, orange and green, respectively.

Uploaded toxicogenomics signatures correspond to tab-delimited text files composed of Entrez gene identifiers (IDs) (25). Prior to running available modules from the workspace, these IDs are supplemented with other related information such as gene symbols, gene descriptions, taxonomy IDs, and HomoloGene IDs (26) using the “gene_info” and “homologene.data” files provided by the NCBI website (<https://www.ncbi.nlm.nih.gov>).

3 Results and discussion

3.1 TOXsIgN overview and current content

TOXsIgN information is organized thanks to a four-layer architecture (Project > Study > Assay > Signature) to which unique trackable identifiers are associated and can be reported in submitted manuscripts, similarly to raw data identifiers from GEO or ArrayExpress (23, 24). Scientists are thus supposed to submit **projects** (each receiving an identifier with the “TSP” prefix) which are subdivided into **studies** (or sub-projects, “TSE” prefix) addressing specific questions and describing experimental **assays** (“TST” prefix) from which specific outcomes are extracted in the form of toxicogenomics **signatures** (“TSS” prefix), i.e. the set of over- and under-expressed genes in the corresponding assays. Of note is that these assays can be performed directly in exposed individuals or in their descendants, and can consist of an exposure to a single environmental factor or a combination of several ones. This definition underlines several unique features of TOXsIgN such as the fact that it is compatible with: *i*) transgenerational studies; *ii*) chemical mixture studies – by default each assay is a mixture of at least one environmental factor; *iii*) several kinds of environmental factors including chemical (e.g. pesticides, plasticizers, drugs, endocrine disruptors), physical (e.g. radiations, temperature) or biological (e.g. pathogens, parasites) factors; and, *iv*) gene sets resulting from several kinds of (transcript-/prote-/epigen-)omics experiments. Importantly, TOXsIgN allows scientists to submit and describe other outcomes than toxicogenomics signatures, such as physiological signatures (e.g. association with a specific phenotype) and molecular signatures (e.g. change in the concentration of a specific hormone) for both interventional (participants undergo some kind of treatment in order to evaluate its impact)

and/or observational studies (individuals for which different outcomes are measured) (27). While growing in terms of heterogeneous signatures, this other feature of TOXsIgN will contribute to break down barriers between different fields of environmental sciences which still remain poorly connected to each other.

Currently, the TOXsIgN repository includes 754 projects for 911 studies that deal with more than 450 compounds performed in human, rat, mouse and drosophila (supplementary table 1 and supplementary information). Together these experimental assays include 8491 toxicogenomics signatures that were extracted from 32688 microarray experiments using 10 different technologies, including data from two major toxicogenomics resources, DrugMatrix and the Toxicogenomics Project-Genomics Assisted Toxicity Evaluation System (Open TG-GATES) (4, 5). The former initiated by the National Institute of Environmental Sciences (NIES, USA) aims at studying the transcriptional response in the rat after exposure to 376 compounds in five different tissues (at multiple doses and multiple exposure times). The latter is a collaborative project between the National Institute of Biomedical Innovation (NIB), the National Institute for Environmental Studies (NIES) and about fifteen pharmaceutical companies intending to study 150 chemicals and their transcriptional responses (at multiple doses, multiple exposure times) in two rat tissues (liver and kidney). Our repository also hosts 326 physiological and molecular signatures from four interventional studies and four observational studies (supplementary table 1). In the next future, toxicogenomics signatures from other research programs, such as CMAP, CEBS and diXa (6, 17, 19), will also feed TOXsIgN.

3.2 Signature submission and access

In order to make the submission procedure quick and easy, all required information has to be recorded by investigators in a dedicated Excel template that embeds one tab for each layer (Project, Study, Assay and Signature). This document integrates a dozen of landmark controlled vocabularies allowing scientists to precisely describe their toxicogenomics studies and corresponding outcomes using ontologies, as recommended (28, 29) (supplementary table 2). Once uploaded, a first evaluation of the Excel template is performed by the TOXsIgN webserver to point out: *i)* “critical errors” for essential information that would not be properly filled (such as the project title) and would therefore prevent the project upload; *ii)* “warnings” for important but not essential missing information (such as a PubMed identifier); and, *iii)* “information” for any other data not appropriately filled (such as additional). If no “critical errors” are detected by the system, the user is next invited to upload the associated toxicogenomics signatures. Each signature consists of three one-column text files corresponding to: *i)* all interrogated genes; *ii)* significantly over-expressed genes; and, *iii)* under-expressed genes. Because reliable and consistent identifier conversion is a complex problem, toxicogenomics signatures should only be composed of Entrez Gene IDs thanks to up-to-date resources (30, 31). When using Affymetrix GeneChip technologies, users are highly recommended to normalize their raw data (CEL files) using the Brainarray custom Chip Description Files (CDF) so that intensity values are not summarized for each probeset but directly for each Entrez gene ID (32) (c.f. supplementary information).

By default, each submitted project and related signatures are tagged with a “private” status meaning that only authorized users (the owner but also co-authors) can access the uploaded data. At this stage, information can still be modified by simply uploading an updated version of the Excel template. For each project, a button is available on the web interface to request the TOXsIgN administrators to switch from a “private” to a “public” status. If some “warnings” are still detected this demand is then rejected. The TOXsIgN administrators will then help the investigators to make the necessary modifications in order to activate the “public” status. Full instructions and examples of the submission procedure are provided in the tutorial section of the TOXsIgN website.

A powerful search engine is implemented to access all “public” information within TOXsIgN. Users can thus interrogate the database based on many distinct fields such as environmental factors, organisms, tissues or technologies. They can also easily make more advanced queries thanks to the intensive use of ontologies to describe toxicogenomics signatures. For instance, an investigator can retrieve 33 toxicogenomics signatures for which the expression of Cyp3a18 (cytochrome P450, family 3, subfamily a, polypeptide 18; a gene encoding a member of the cytochrome P450 superfamily highly expressed in the liver (33)) is negatively affected in liver after exposure to hydroxyl steroid compounds (Figure 1).

In addition to toxicogenomics signatures archived in TOXsIgN, other data are also made available in a dedicated web page accessible through the “Download” tab on the main interface. It notably includes the last release of the Excel template document for uploading signatures in the repository and all ontologies (OBO files) used in TOXsIgN.

3.3. Inferring adverse effects from toxicogenomics signatures

Predictive toxicology approaches that are based on toxicogenomics data intend to evaluate compounds' toxicity using altered gene expression as an endpoint. In 2004, Steiner and colleagues established the proof-of-concept that close toxicogenomics signatures imply close adverse effects (8). In this study, the authors used support vector machines (SVMs) to classify hepatotoxic and non-hepatotoxic chemicals based on transcriptomic data. Although parameters were likely overfitted due to the small number of compounds, this approach properly predicted hepatotoxic effects for 90% of known hepatotoxicants.

To illustrate the usefulness of archiving massive toxicogenomics signatures in a public repository we tried demonstrate the hypothesis formulated by Steiner and colleagues at a higher scale (supplementary information). A subset of the current TOXsIgN repository including 3022 toxicogenomics signatures from 5 rat tissues after exposure to 410 toxicants was used and the degree of linear correlation between basic toxicogenomics distances (Pearson's correlation and Euclidean distance calculated between two toxicogenomics signatures) and the toxicological distance (concordance, i.e. the number of shared adverse effects between two toxicants) was evaluated. We first confirmed a highly significant association between toxicogenomics and toxicological distances, particularly by using the Pearson's correlation for the toxicogenomics distance ($r^2 = 0.051$, $P \leq 0.0$) (Figure 2, panels A-B). This result thus confirms the hypothesis that a significant correlation between gene expression profiles also implies close toxicities. Strikingly, we observed a slightly better linear correlation by using a discretized

expression data ($r^2 = 0.056$, $P \leq 0.0$; panels C-D), i.e. an expression matrix in which fold-change information were simplified to only three distinct status (c.f. supplementary information): 1, over-expressed genes after exposure; -1, under-expressed; and, 0, no differential expression. This strengthens the idea that toxicogenomics signatures can be archived as simplified text files, thus simplifying their submission, without penalizing their predictive potential.

3.4 The signature enrichment analysis module for comparing toxicogenomics signatures

The core feature of the TOXsIgN workspace is to provide bioinformatics and biostatistics modules allowing retrieval, analysis and comparison of the toxicogenomics signatures uploaded in the repository. The cross-species and cross-technology compatibility of this workspace relies on the fact that toxicogenomics signatures are first converted into HomoloGene IDs prior to being used by the different tools. On the web page dedicated to each toxicogenomics signature, a “Save in workspace” button allows investigators to transfer the corresponding signature into the workspace for further analyzes, while all available modules can be accessed through the “Tools” tab from the main interface.

In addition to the search engine and other conversion tools, three other modules are already available via the interface:

- i) *Signature comparison*. This module embeds an interactive Venn diagram viewer, called jvenn, to easily compare up to six selected toxicogenomics signatures (34).
- ii) *Functional enrichment analysis*. This tool allows user to explore the mechanisms of toxicity of a given environmental factor by identifying biological processes, molecular functions, cellular components and phenotypes associated to its toxicogenomics signature. Briefly, it determines the significance of the resulting overlaps by using the hypergeometric probability.
- iii) *Signature enrichment analysis*. This module aims at identifying closely-related toxicogenomics signatures in the repository as compared to a selected one. Similar to *Functional enrichment analysis*, the hypergeometric probability is used to determine the significance of the overlaps. It also includes a distance matrix calculation using the Euclidean distance and the Pearson's correlation to discriminate the closely-related toxicogenomics signatures.

To illustrate the relevance of these modules, we uploaded in the workspace a toxicogenomics signature corresponding to 381 over-expressed and 494 under-expressed genes in the rat liver after exposure to diethylstilbestrol (a well-known estrogenic endocrine-disrupting chemical (35, 36)) for 5 days, at 2.8 mg/kg (TOXsIgN signature identifier: TSS230). We next used the *Signature enrichment analysis* module using default parameters, which took about one minute to complete the corresponding job: consistently, three of the top-10 toxicogenomics signatures includes experiments performed in the rat liver after exposure to the same exact compound but at different doses (three experimental conditions) (Figure 3). The seven other signatures correspond to other well-known estrogenic compounds sharing similar mechanisms of action with DES (such as ethinylestradiol, estriol, mestranol and β -estradiol) (37, 38), yet confirming the proof-of-concept established by Steiner and colleagues (8).

In the near future, the TOXsIgN workspace will also provide advanced tools leaning on the archived toxicogenomics signatures that will help investigators predict and prioritize the toxicological effects of environmental factors relevant to their specific interests.

4 Conclusion and perspectives

Our goal when designing TOXsIgN was to develop a new cross-species repository allowing scientists to easily submit their own published toxicogenomics which we believe to be of very high quality data since they are evaluated by experts during the peer-review process. By definition, TOXsIgN is neither intended to archive raw data such as GEO and ArrayExpress (23, 24), nor to replace existing toxicological databases(16, 17, 19), but rather to complement these resources by acting as a distribution hub. We believe TOXsIgN could constitute a new alternative for toxicological data storage, accessibility and reusability and could thus improve toxicological experiment reproducibility. The success of such repository obviously resides on the scientists' willingness to make their data FAIR, but also on scientific journal's editors. Indeed, making raw data available to the community could be considered insufficient given the demand of the community to directly and easily access process interpreted data. We propose that the same effort could be done to encourage scientists to upload toxicogenomics and toxicological signatures in TOXsIgN prior to submitting a manuscript for publication. On our side, we already integrating additional toxicogenomics signatures from other major toxicogenomics projects such as CMAP, CEBS and diXa (6, 17, 19). Likewise, we also plan to make TOXsIgN compatible with other kind of environmental factors, such as physical and biological factors. Altogether, we expect that this new resource could significantly contribute to risk assessment.

In addition to serving as a public repository, TOXsIgN also intends to become a warehouse for toxicogenomics and predictive toxicology tools. Its modular design facilitates the implementation of additional bio-informatics modules leaning on the deposited toxicogenomics signatures that will help investigators analyze and predict adverse effects of environmental factors relevant to their specific interests. Several changes are already planned: First, prediction and prioritization systems for chemical toxicity currently under development in our lab will be made available. We will also implement a module to automatically extract toxicogenomics signatures from raw data by using our own workflow. Another important aspect would be to add social features in TOXsIgN allowing several investigators to work on the same data. Together these efforts will improve TOXsIgN's utility, making a front-line resource relevant to a large audience, including toxicologists, biologists, epidemiologists and environmental sciences in general.

Acknowledgements

We thank the GenOuest bioinformatics facility for hosting the software as well as all members of the Institute for research in Health, Environment and Work for stimulating discussions. We also thank the Institut national de la santé et de la recherche médicale (Inserm), the Centre national de la recherche

scientifique (CNRS), l'Université de Rennes 1 and the French School of Public Health (EHESP) for supporting this work.

Funding

TOXsIgN is supported, built and maintained by the Research Institute for Environmental and Occupational Health (IRSET), the French School of Public Health (EHESP) and the GenOuest Bioinformatics core facility. This work was supported by the Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail [ANSES n°EST-13-081 to F.C.] ; the Fondation pour la recherche médicale [FRM n°DBI20131228558 to F.C.], and the European Union [FEDER to F.C.]. Funding for open access charge: Institut national de la santé et de la recherche médicale (Inserm) and l'Université de Rennes 1.

Conflict of interest: None declared

References

1. Tweedale,A.C. (2017) The inadequacies of pre-market chemical risk assessment's toxicity studies- the implications. *J. Appl. Toxicol.*, **37**, 92–104.
2. European Commission REGULATION (EC) No 1907/2006 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 18 December 2006 concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), establishing a European Chemicals Agency, amending Directive 1999/4.
3. RUSSELL,W.M.S. and BURCH,R.L. (1959) The principles of humane experimental technique. *Princ. Hum. Exp. Tech.*
4. Igarashi,Y., Nakatsu,N., Yamashita,T., Ono,A., Ohno,Y., Urushidani,T. and Yamada,H. (2015) Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.*, **43**, D921-7.
5. Ganter,B., Snyder,R.D., Halbert,D.N. and Lee,M.D. (2006) Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. *Pharmacogenomics*, **7**, 1025–44.
6. Lamb,J., Crawford,E.D., Peck,D., Modell,J.W., Blat,I.C., Wrobel,M.J., Lerner,J., Brunet,J.-P., Subramanian,A., Ross,K.N., *et al.* (2006) The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science (80-.)*, **313**, 1929–1935.
7. Prathipati,P. and Mizuguchi,K. (2016) Systems Biology Approaches to a Rational Drug Discovery Paradigm. *Curr. Top. Med. Chem.*, **16**, 1009–25.
8. Steiner,G., Suter,L., Boess,F., Gasser,R., de Vera,M.C., Albertini,S. and Ruepp,S. (2004) Discriminating different classes of toxicants by transcript profiling. *Environ. Health Perspect.*, **112**,

9. Kavlock,R.J., Dix,D.J., Houck,K.A., Judson,R.S., Martin,M.T. and Richard,A.M. (2007) ToxCast TM : Developing predictive signatures for chemical toxicity.
10. Must try harder (2012) *Nature*, **483**, 509–509.
11. Miller,G.W. (2014) Improving Reproducibility in Toxicology. *Toxicol. Sci.*, **139**, 1–3.
12. George,B.J., Sobus,J.R., Phelps,L.P., Rashleigh,B., Simmons,J.E., Hines,R.N. and Community of Practice for Statistics Guidance Documents Working Groups (2015) Raising the bar for reproducible science at the U.S. Environmental Protection Agency Office of Research and Development. *Toxicol. Sci.*, **145**, 16–22.
13. Poland,C.A., Miller,M.R., Duffin,R. and Cassee,F. (2014) The elephant in the room: reproducibility in toxicology. *Part. Fibre Toxicol.*, **11**, 42.
14. Collins,F.S. and Tabak,L.A. (2014) Policy: NIH plans to enhance reproducibility. *Nature*, **505**, 612–3.
15. Wilkinson,M.D., Dumontier,M., Aalbersberg,I.J., Appleton,G., Axton,M., Baak,A., Blomberg,N., Boiten,J.-W., da Silva Santos,L.B., Bourne,P.E., *et al.* (2016) The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data*, **3**, 160018.
16. Davis,A.P., Grondin,C.J., Lennon-Hopkins,K., Saraceni-Richards,C., Sciaky,D., King,B.L., Wieggers,T.C. and Mattingly,C.J. (2015) The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.*, **43**, D914–20.
17. Hendrickx,D.M., Aerts,H.J.W.L., Caiment,F., Clark,D., Ebbels,T.M.D., Evelo,C.T., Gmuender,H., Hebels,D.G.A.J., Herwig,R., Hescheler,J., *et al.* (2015) diXa: a data infrastructure for chemical safety assessment. *Bioinformatics*, **31**, 1505–1507.
18. Hardt,C., Beber,M.E., Rasche,A., Kamburov,A., Hebels,D.G., Kleinjans,J.C. and Herwig,R. (2016) ToxDB: pathway-level interpretation of drug-treatment data. *Database (Oxford)*, **2016**, baw052.
19. Lea,I.A., Gong,H., Paleja,A., Rashid,A. and Fostel,J. (2017) CEBS: a comprehensive annotated database of toxicological data. *Nucleic Acids Res.*, **45**, D964–D971.
20. Roider,H.G., Pavlova,N., Kirov,I., Slavov,S., Slavov,T., Uzunov,Z. and Weiss,B. (2014) Drug2Gene: an exhaustive resource to explore effectively the drug-target relation network. *BMC Bioinformatics*, **15**, 68.
21. Miller,G.W. (2015) Data sharing in toxicology: beyond show and tell. *Toxicol. Sci.*, **143**, 3–5.
22. Richard,A.M., Judson,R.S., Houck,K.A., Grulke,C.M., Volarath,P., Thillainadarajah,I., Yang,C., Rathman,J., Martin,M.T., Wambaugh,J.F., *et al.* (2016) ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.*, **29**, 1225–1251.
23. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M., *et al.* (2013) NCBI GEO: archive for functional genomics

- data sets--update. *Nucleic Acids Res.*, **41**, D991-5.
24. Kolesnikov,N., Hastings,E., Keays,M., Melnichuk,O., Tang,Y.A., Williams,E., Dylag,M., Kurbatova,N., Brandizi,M., Burdett,T., *et al.* (2015) ArrayExpress update--simplifying data submissions. *Nucleic Acids Res.*, **43**, D1113-6.
 25. Maglott,D., Ostell,J., Pruitt,K.D. and Tatusova,T. (2011) Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.*, **39**, D52-7.
 26. NCBI Resource Coordinators (2016) """". *Nucleic Acids Res.*, **44**, D7–D19.
 27. Thiese,M.S. (2014) Observational and interventional study design types; an overview. *Biochem. medica*, **24**, 199–210.
 28. Hardy,B., Apic,G., Carthew,P., Clark,D., Cook,D., Dix,I., Escher,S., Hastings,J., Heard,D.J., Jeliaskova,N., *et al.* (2012) Toxicology ontology perspectives. *ALTEX*, **29**, 139–56.
 29. Smith,B., Ashburner,M., Rosse,C., Bard,J., Bug,W., Ceusters,W., Goldberg,L.J., Eilbeck,K., Ireland,A., Mungall,C.J., *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.*, **25**, 1251–5.
 30. Huang,D.W., Sherman,B.T., Stephens,R., Baseler,M.W., Lane,H.C. and Lempicki,R.A. (2008) DAVID gene ID conversion tool. *Bioinformatics*, **2**, 428–30.
 31. Mudunuri,U., Che,A., Yi,M. and Stephens,R.M. (2009) bioDBnet: the biological database network. *Bioinformatics*, **25**, 555–6.
 32. Dai,M., Wang,P., Boyd,A.D., Kostov,G., Athey,B., Jones,E.G., Bunney,W.E., Myers,R.M., Speed,T.P., Akil,H., *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.
 33. Nagata,K., Murayama,N., Miyata,M., Shimada,M., Urahashi,A., Yamazoe,Y. and Kato,R. (1996) Isolation and characterization of a new rat P450 (CYP3A18) cDNA encoding P450(6)beta-2 catalyzing testosterone 6 beta- and 16 alpha-hydroxylations. *Pharmacogenetics*, **6**, 103–11.
 34. Bardou,P., Mariette,J., Escudié,F., Djemiel,C. and Klopp,C. (2014) jvenn: an interactive Venn diagram viewer. *BMC Bioinformatics*, **15**, 293.
 35. Korach,K.S. and McLachlan,J.A. (1985) The role of the estrogen receptor in diethylstilbestrol toxicity. *Arch. Toxicol. Suppl.*, **8**, 33–42.
 36. Li,Y., Hamilton,K.J., Lai,A.Y., Burns,K.A., Li,L., Wade,P.A. and Korach,K.S. (2013) Diethylstilbestrol (DES)-Stimulated Hormonal Toxicity is Mediated by ER α Alteration of Target Gene Methylation Patterns and Epigenetic Modifiers (DNMT3A, MBD2, and HDAC2) in the Mouse Seminal Vesicle. *Environ. Health Perspect.*, **122**, 262–8.
 37. MUECHLER,E.K. and KOHLER,D. (1980) Properties of the Estrogen Receptor in the Human Oviduct and Its Interaction with Ethinylestradiol and Mestranol in Vitro*. *J. Clin. Endocrinol. Metab.*, **51**, 962–967.

38. Simpson,E. and Santen,R.J. (2015) Celebrating 75 years of oestradiol. *J. Mol. Endocrinol.*, **55**, T1–T20.

Figures

TOXsIgN

[Home](#) [Browse](#) [Download](#) [Tools](#) [Q Search](#) [Running jobs](#) [Log in](#) [Register](#)

[Advanced search](#)

Advanced search

Search parameters

AND

Select a field

Add

AND - (.type:signatures AND genes_down:"252931")

AND - (.type:signatures AND tissue:"Liver")

AND - (.type:signatures AND tags:"steroid")

Search

Search results

Projets (No Result)

Studies (No Result)

Signatures (33)

TSS203 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (10 mgkg, 5 days) in the rat

[\(view\)](#)

Type : Genomic

Organism : Rattus norvegicus

Project : TSP62

Study : TSE80

Assay : TSA203

TSS208 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (10 mgkg, 3 days) in the rat

[\(view\)](#)

Type : Genomic

Organism : Rattus norvegicus

Project : TSP62

Study : TSE80

Assay : TSA208

TSS206 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (1480 mgkg, 3 days) in the rat

[\(view\)](#)

Type : Genomic

Organism : Rattus norvegicus

Project : TSP62

Study : TSE80

Assay : TSA206

TSS205 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (10 mgkg, .25 days) in the rat

[\(view\)](#)

Type : Genomic

Organism : Rattus norvegicus

Project : TSP62

Study : TSE80

Assay : TSA205

Figure 1- The TOXsIgN advance search engine. Results for toxicogenomics signatures in which the expression of *Cyp3a18* (Gene id: 252931) is *downregulated* in the rat *liver* after exposure to *hydroxyl steroid compounds*. From top to bottom, “search parameters”, “full query” and “results table” are displayed. “Search parameters” allow users to describe their request among projects, studies, assays or signatures. For each request, the ElasticSearch query used to retrieve information is displayed on the top of the “results table”. The “results table” is a three-tab panel displaying result sorted according to projects, studies or signatures and providing a link towards the corresponding page.

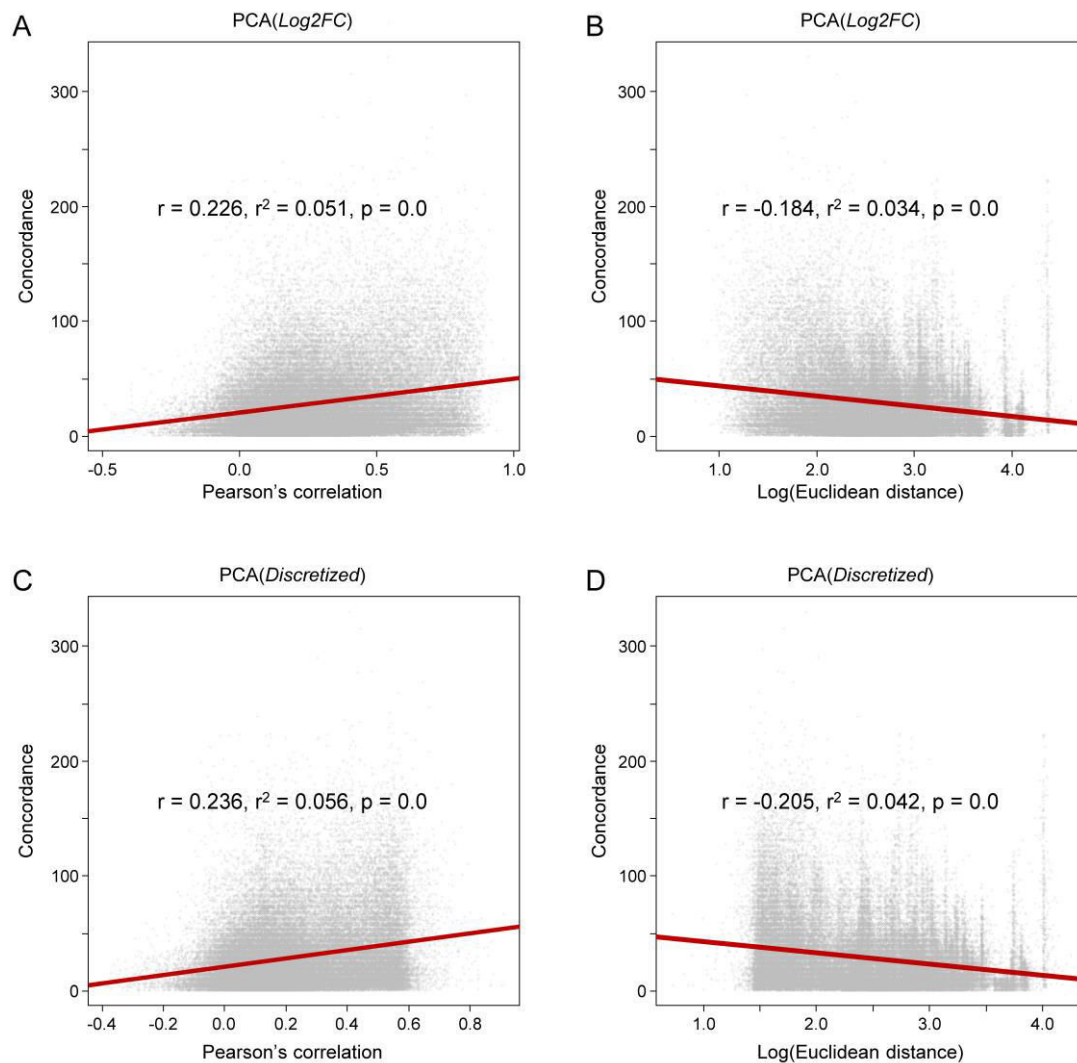


Figure 2- Correlation between toxicogenomics distances and toxicological distance. The linear correlation between toxicogenomics distances (Pearson's correlation and Euclidean distance calculated between two toxicogenomics signatures) and the toxicological distances (concordance, i.e. the number of shared adverse effects between two toxicants) was evaluated. Panel A-B show a significant association between toxicogenomics and toxicological distances using the Pearson's correlation for the toxicogenomics distance on the log-fold-change matrix ($r^2 = 0.051$, $P \leq 0.0$). Panels C-D demonstrate better linear correlation by using a discretized expression matrix ($r^2 = 0.056$, $P \leq 0.0$), i.e. expression data where fold-change information were discretized in only three distinct status (1, over-expressed after exposure; -1, under-expressed; and, 0, no differential expression).

Signature	r	R	n	N	Ratio	Pvalue	Zscore	Euclidean distance	Correlation distance
<input type="text"/>									
TSS203 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (10 mgkg, 5 days) in the rat	637	842	907	10770	57 %	0	73.1634644491449	21.7944947177034	0.729101755005883
TSS229 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to diethylstilbestrol (280 mgkg, 3 days) in the rat	604	842	943	10770	51 %	0	67.336760881398	24.0208242989286	0.680260240200066
TSS208 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (10 mgkg, 3 days) in the rat	542	842	770	10770	51 %	0	67.1180092422705	22.9782505861521	0.67384691254637
TSS226 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to diethylstilbestrol (280 mgkg, 5 days) in the rat	649	842	1104	10770	50 %	0	66.5851006025112	25.4558441227157	0.674621819620628
TSS228 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to diethylstilbestrol (2.8 mgkg, 3 days) in the rat	520	842	743	10770	49 %	0	65.4178016993892	23.3452350598575	0.66040973201856
TSS1179 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to beta-estradiol (150 mgkg, 5 days) in the rat	565	842	1053	10770	42 %	0	58.3300260758659	27.6947648482525	0.604443584977887
TSS1183 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to beta-estradiol (150 mgkg, 3 days) in the rat	449	842	721	10770	40 %	0	56.3862533359308	25.8263431402899	0.576911914376368
TSS783 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to mestranol (250 mgkg, 5 days) in the rat	518	842	960	10770	40 %	0	55.7953723188752	27.712812921102	0.576529708714113
TSS1117 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to estriol (313 mgkg, 5 days) in the rat	556	842	1164	10770	38 %	0	53.7551493006238	29.8998327754521	0.563817014922414
TSS207 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (1480 mgkg, 5 days) in the rat	620	842	1519	10770	36 %	0	51.6883600069076	33.5111921602321	0.549224727537796

Figure 3- Signature enrichment analysis for Diethylstilbestrol. Diethylstilbestrol (DES) toxicogenomics signature (2.8 mg/kg, 5 days, rat liver) was compared to all toxicogenomics signatures indexed in TOXsIgN using the Signature enrichment analysis tool. This signature shows significant enrichment with three other toxicogenomics signatures obtained after exposure to DES. Other estrogenic chemicals sharing similar mode(s) of action such as ethinylestradiol, estriol, mestranol and β -estradiol are found to significantly overlap the selected signature. N is the total number of HomoleGene IDs measured, R is the total number of HomoleGene IDs meeting the criterion, n is the total number of HomoleGene IDs in the selected signature, r is the number of HomoleGene IDs meeting the criterion in this signature, while the ratio corresponds to $r/\text{union}(n, R)$. *P-value* is obtained using the hypergeometric probability law adjusted for multiple testing. Euclidean distance and Pearson's correlation were also calculated to estimate the distance between two toxicogenomics signatures.

Supplementary information

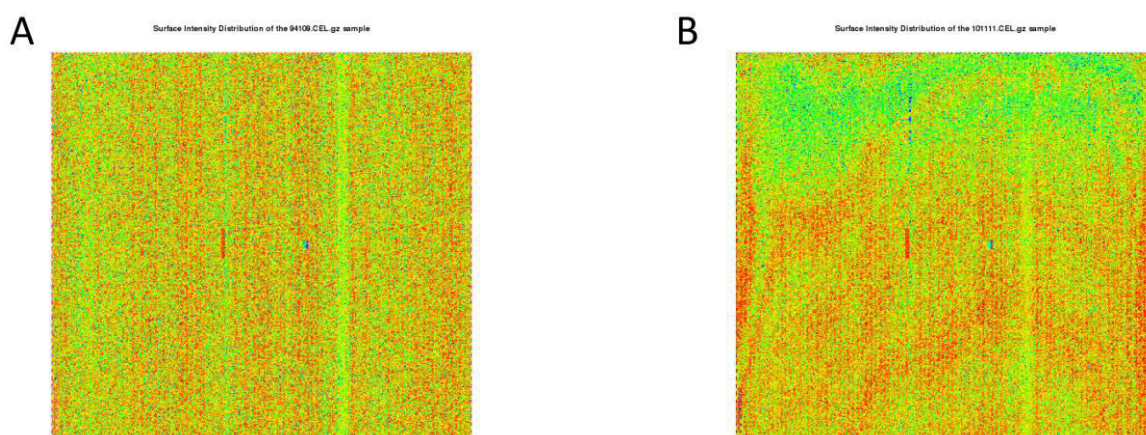
1. Current toxicogenomics signatures hosted in TOXsIgN

1.1. A massive dataset of toxicogenomics microarray studies

In July 2017, the TOXsIgN repository hosted toxicogenomics studies assembled from Open TG-GATEs (1) and DrugMatrix (2) as well as from the Gene Expression Omnibus public repository (3). Together we assembled a transcriptomic dataset composed of 32688 raw data files (Affymetrix CEL files) from 9 technologies corresponding to 8491 experimental conditions in 4 species (supplementary table 1). An experimental condition is defined as an experiment performed on one sample (tissue or cell line) exposed at one chemical at one dose and one exposure time.

1.2. Quality control and raw data preprocessing

The data in the resulting CEL files were processed using the *affy* package available in Bioconductor (4). Briefly, the RMA method was used for data normalization, background correction and summarization by using appropriate custom CDF files (version 20.0.0, ENTREZG) available at the Brainarray website (<http://brainarray.mbni.med.umich.edu>) (5). Experimental conditions and their related exposed and control samples were preprocessed separately. The data were quality controlled by box plots, Kullback-Leibler distance matrix and surface intensity distribution as previously described in (6). Among the 32688 CEL files, 68 failed to pass the quality control procedure mostly because damages or scratches exceeded 20% of the array surface (supplementary figure 1).

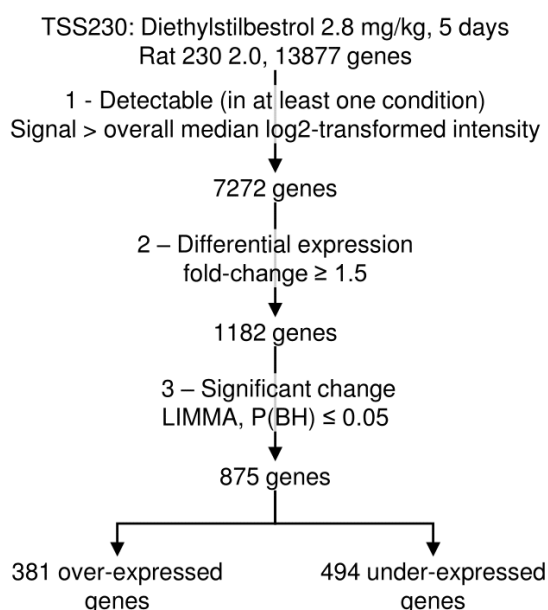


Supplementary figure 1. Quality control using surface intensity distribution. Panel A shows the surface intensity distribution of a good-quality CEL file. Panel B shows the surface intensity distribution of a CEL file that failed to pass this test as scratches exceeded 20% of its array surface.

1.3. Statistical data analysis and extraction of toxicogenomics signatures

Genes showing a differential expression among exposed and control samples of a given experimental condition were identified by applying three filtering steps (supplementary figure 2). First detectable genes with at least one signal above the background expression cutoff corresponding to the overall median log2-transformed intensity were selected. Next, the subset of detectable genes showing highly variable signals between exposed and control samples was filtered (fold-change ≥ 1.5). Finally, genes showing a significant change among the samples ($P \leq 0.05$) were filtered by using the statistical test implemented in the LIMMA (linear models for microarray data) package (7). *P-values* were corrected for multiple testing by using the false discovery rate (FDR) adjustment method. The set of genes significantly over- and under-expressed after exposure to a chemical (by comparing exposed and control samples) that results from this statistical filtration constitutes the toxicogenomics signature of the corresponding experimental condition.

This strategy allowed us to identify one toxicogenomics signature for the 8491 experimental conditions. Among them, 3719 experimental conditions modified the expression of at least 10 genes and 3734 were not associated with significant and detectable transcriptional changes after exposure.



Supplementary figure 2. Toxicogenomics signature workflow. A schematic diagram of the strategy used to identify the toxicogenomics signature of a chemical compound. The diethylstilbestrol signature obtained from Rat 230 2.0 microarray was selected (TOXsIgN signature identifier: TSS230). This microarray interrogates 13877 genes. First detectable genes with at least one signal above the background expression cutoff corresponding to the overall median log2-transformed intensity were selected. 7272 genes passed this test. Next, detectable genes showing highly variable signals between exposed and control samples was filtered (fold-change ≥ 1.5). From the previously 7272 selected genes, only 1182 succeeded to pass the test. Finally, genes showing a significant change among the samples ($P \leq 0.05$) were filtered by using the statistical test implemented in the LIMMA (linear models for microarray data) package. *P-values* were corrected for multiple testing by using the false discovery rate (FDR) adjustment method. From the 1182 genes, 875 was selected including 381 over-expressed genes and 494 under-expressed genes.

1.4. Assembly of massive toxicogenomics signature matrices to validate the hypothesis formulated by Steiner and colleagues

This section focus on the Steiner's proof-of-concept (8) suggesting that close toxicogenomics signatures may have close mechanisms of toxicity.

Among the 8491 toxicogenomics signatures deposited in the TOXsIgN repository we assembled a subset of 7505 signatures from experiments performed with the exact same technology, the Affymetrix Rat GeneChip 230 2.0 (enabling to interrogate the expression level of 13877 genes). Next, we selected 452 chemicals associated with 7465 toxicogenomics for which known adverse effects were described in the literature and reported by the Comparative Toxicogenomics Database using a disease ontology (CTD) (9).

A		<i>Log2FC</i> 11433 genes				
		A	G ₁	G ₂	G ₃	... G ₁₁₄₃₃
3022 conditions	C ₁	-0.4	0	0.6	-1.9	
	C ₂	0	0	1.7	-0.7	
	C ₃	0	1.5	-2.1	0	
	...					
	C ₃₀₂₂	-2.4	1.6	0	1	

B		<i>Discretized</i> 11433 genes				
		G ₁	G ₂	G ₃	... G ₁₁₄₃₃	
3022 conditions	C ₁	0	0	1	-1	
	C ₂	0	0	1	-1	
	C ₃	0	1	-1	0	
	...					
	C ₃₀₂₂	-1	1	0	1	

C		<i>PCA(Log2FC)</i> 2741 Components				
		D ₁	D ₂	D ₃	... D ₂₇₄₁	
3022 conditions	C ₁	-1.2	0.4	-5.6	-1.9	
	C ₂	-0.5	1.9	-6.2	-0.7	
	C ₃	0.7	-0.9	0.9	0.7	
	...					
	C ₃₀₂₂	-7.4	6.6	0.1	3.1	

D		<i>PCA(Discretized)</i> 2670 Components				
		D ₁	D ₂	D ₃	... D ₂₆₇₀	
3022 conditions	C ₁	0.2	0.9	-4	2.2	
	C ₂	-0.1	-0.3	0.7	-1.3	
	C ₃	-3.2	-1.8	2.5	0.9	
	...					
	C ₃₀₂₂	7.2	2.4	-0.8	-1.6	

Supplementary figure 3. Toxicogenomics matrices. Four distinct types of toxicogenomics matrices were produced. Panel A shows the initial matrix, called *Log2FC*, in which each line represents experimental conditions and each column represents a gene. Log2-transformed fold-change values are reported in each cell of this matrix. In the second matrix, called *discretized*, expression changes were discretized by using the following rules: genes over- and under-expressed were associated with a “1” and “-1” status and genes not differentially expressed after exposure with a “0” status (panel B). A principal component analysis of both matrices *Log2FC* and *discretized* were then performed resulting in to other matrices called *PCA(Log2FC)* and *PCA(discretized)* (Panels C and D).

Two toxicogenomics matrices were assembled in which rows correspond to 7465 signatures and columns to 13877 genes (supplementary figure 3). In the first matrix, called *log2fc* matrix, a log2-transformed fold-change value between exposed and control samples is reported for each signature and each gene. In the second matrix, called *discretized* matrix, expression change information were

discretized using only three status: “1” and “-1”, over- and under-expressed genes after exposure (c.f. **1.3. Statistical data analysis and extraction of toxicogenomics signatures**); and, “0”, no significant differential changes detected. Next, we reduce these two matrices by selecting: toxicogenomics signatures composed of more than 10 genes showing an altered expression after exposure; and, genes showing a significant change in expression in at least one toxicogenomics signature. Both resulting toxicogenomics matrices encompassed 3022 toxicogenomics signatures (rows) from 410 toxicants and 11433 genes (columns).

Finally, both matrices were subjected to principal component analyses using the *FactoMineR* package (10) implemented in R (11) in which genes were used as variables. This technique is particularly suited to reduce massive data to low-dimensional space only composed of orthogonal components maximizing the variance. Only the top-ranked components explaining the greatest amount of the variance by using the *Estim_ncp_PCA* function implemented *FactoMineR* were selected for further analyses (2670 and 2741 components for the *log2fc* and *discretized* matrices, respectively) resulting in two other matrices called *PCA(log2fc)* and *PCA(discretized)*.

3. References

1. Y. Igarashi *et al.*, Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res.* **43**, D921-7 (2015).
2. B. Ganter, R. D. Snyder, D. N. Halbert, M. D. Lee, Toxicogenomics in drug discovery and development: mechanistic analysis of compound/class-dependent effects using the DrugMatrix database. *Pharmacogenomics*. **7**, 1025–44 (2006).
3. NCBI Resource Coordinators, """". *Nucleic Acids Res.* **44**, D7–D19 (2016).
4. W. Huber *et al.*, Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods*. **12**, 115–21 (2015).
5. M. Dai *et al.*, Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175–e175 (2005).
6. F. Chalmel *et al.*, The conserved transcriptome in human and rodent male gametogenesis. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 8346–51 (2007).
7. M. E. Ritchie *et al.*, limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
8. G. Steiner *et al.*, Discriminating different classes of toxicants by transcript profiling. *Environ. Health Perspect.* **112** (2004), pp. 1236–48.
9. A. P. Davis *et al.*, The Comparative Toxicogenomics Database's 10th year anniversary: update 2015. *Nucleic Acids Res.* **43**, D914-20 (2015).
10. S. Lê, J. Josse, F. Husson, FactoMineR: An R Package for Multivariate Analysis. *J. Stat. Softw.* **25** (2008), pp. 1–18.
11. R core Team, *R: a language and environment for statistical computing* (2013).

Supplementary table 1. TOXsIgN repository content. Currently TOXsIgN hosts 8491 toxicogenomics signatures and 326 molecular and physiological signatures of more than 450 chemicals on 17 different tissues or cell lines in five species.

GSE	PubMed ID	Species	Study type	Tissue / Cell	n° of Chemicals	Signature type	Signature	Technology
GSE10748	19212759	Rat	Interventional	Brain	1	Genomic	5	GPL1355
GSE57800	25058030	Rat	Interventional	Heart	88	Genomic	206	GPL1355
GSE19366	21865292	Rat	Interventional	Kidney	3	Genomic	24	GPL1355
GSE57811	25058030	Rat	Interventional	Kidney	139	Genomic	365	GPL1355
TG-GATEs	25313160	Rat	Interventional	Kidney	150	Genomic	1363	GPL1355
GSE57815	25058030	Rat	Interventional	Liver	200	Genomic	654	GPL1355
TG-GATEs	25313160	Rat	Interventional	Liver	150	Genomic	4776	GPL1355
GSE8238	17127748	Rat	Interventional	Ovary	1	Genomic	1	GPL1355
GSE10557	17660231	Rat	Interventional	Ovary	1	Genomic	1	GPL1355
GSE32890	22808131	Rat	Interventional	Ovary	1	Genomic	1	GPL1355
GSE9480	/	Rat	Interventional	Testis	1	Genomic	2	GPL1355
GSE10412	19423681	Rat	Interventional	Testis	3	Genomic	6	GPL1355
GSE10919	18042343	Rat	Interventional	Testis	1	Genomic	3	GPL1355
GSE20245	20566332	Rat	Interventional	Testis	2	Genomic	5	GPL1355
GSE9387	19409404	Rat	Interventional	Primary hepatocyte	5	Genomic	10	GPL1355
GSE10015	20521778	Rat	Interventional	Liver	2	Genomic	24	GPL1355
GSE57816	25058030	Rat	Interventional	Thigh muscle	21	Genomic	42	GPL1355
GSE10093	16972790	Mouse	Interventional	CD8 T cells	1	Genomic	1	GPL339
GSE10408	19409404	Rat	Interventional	Liver	2	Genomic	6	GPL1355
GSE10409	19409404	Rat	Interventional	Liver	3	Genomic	3	GPL1355
GSE10411	19409404	Rat	Interventional	Liver	3	Genomic	6	GPL1355
GSE11695	19737576	Drosophila	Interventional	Whole body	1	Genomic	1	GPL1322
GSE14553	19692669	Human	Interventional	Primary hepatocyte	2	Genomic	27	GPL96
GSE14554	19692669	Rat	Interventional	Primary hepatocyte	2	Genomic	14	GPL85
GSE30861	23221170	Mouse	Interventional	Liver	1	Genomic	6	GPL1261
GSE31540	25448281	Mouse	Interventional	Liver	1	Genomic	9	GPL1261
GSE40117	23393228	Rat	Interventional	Human embryonic	15	Genomic	15	GPL570
GSE44783	23393228	Mouse	Interventional	Liver	13	Genomic	56	GPL1261
GSE48126	24535564, 25448281	Mouse	Interventional	Liver	13	Genomic	15	GPL1261
GSE51969	25448281	Mouse	Interventional	Liver	13	Genomic	12	GPL1261
GSE53082	24830643	Rat	Interventional	Liver	13	Genomic	15	GPL341
GSE53634	25270620	Mouse	Interventional	Primary hepatocyte	13	Genomic	15	GPL1261

GSE72081	27338304	Mouse	Interventional	Hepatocytes	13	Genomic	42	GPL1261
GSE72755	26697389	Mouse	Interventional	Liver	13	Genomic	2	GPL6246
GSE74676	26775027	Rat	Interventional	Hippocampus	13	Genomic	2	GPL1355
GSE10410	19409404	Human	Interventional	primary hepatocytes	13	Genomic	11	GPL570
GSE11869	18936297	Human	Interventional	Ishikawa cells	13	Genomic	12	GPL570
GSE27094	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27095	19159671	Human	Interventional	HK-2	13	Genomic	3	GPL570
GSE27096	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27167	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27168	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27169	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27170	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27182	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27188	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27189	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27190	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27191	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27192	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27196	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27198	19159671	Human	Interventional	HK-2	13	Genomic	2	GPL570
GSE27202	19159671	Human	Interventional	HK-2	13	Genomic	10	GPL570
GSE27204	19159671	Human	Interventional	HK-2	13	Genomic	8	GPL570
GSE27208	19159671	Human	Interventional	HK-2	13	Genomic	9	GPL570
GSE27210	19159671	Human	Interventional	HK-2	13	Genomic	26	GPL570
GSE46909	23824090	Human	Interventional	Jurkat cells	13	Genomic	32	GPL570
GSE50705	24062438	Human	Interventional	MCF-7	13	Genomic	86	GPL570
/	in press	Human	Interventional	Testis	13	Physiologic al	4	/
/	23547264	Chinese hamster	Interventional	Ovary	13	Molecular	122	/
/	24030937	Human	Interventional	Testis	13	Molecular	45	/
/	25681860	Human	Observational	/	13	Physiologic al	38	/
/	26445054	Human	Observational	/	13	Physiologic al	48	/
/	24100206	Human	Observational	/	13	Physiologic al	9	/
/	27740510	Human	Observational	/	13	Physiologic al	60	/

Supplementary table 2. Controlled vocabularies. The TOXsIgN repository encompasses 12 distinct ontologies to structure information such as chemical, tissue or cell line, associated adverse effects. Those controlled vocabularies allow investigators to precisely describe toxicogenomics studies and their corresponding outcomes.

Name	Abbreviation	PMID	TOXsIgN Excel field	Link
National Center for Biotechnology Information (NCBI) Organismal Classification	NCBITAXON	27899561	Species	https://www.ncbi.nlm.nih.gov/taxonomy
Chemical Entities of Biological Interest	CHEBI	23180789	Chemical	https://www.ebi.ac.uk/chebi/
Foundational Model of Anatomy	FMA	14759820	Tissue	http://si.washington.edu/projects/fma
Human Phenotype Ontology	HP	18950739	Phenotype	http://human-phenotype-ontology.github.io/
Phenotypic Quality Ontology	PATO	/	Phenotype	http://www.obofoundry.org/ontology/pato.html
Human Disease Ontology	DOID	25348409	Phenotype	http://www.disease-ontology.org/
Gene Ontology	GO	14681407	Molecular/Cellular phenotype	http://www.geneontology.org/
Cell Ontology	CL	15693950	Cell	https://www.ebi.ac.uk/ols/ontologies/cl
Cell Line Ontology	CLO	25852852	Cell Line	http://www.clo-ontology.org/
BRENDA Tissue and Enzyme Source Ontology	BTO	25378310	Tissue	http://www.brenda-enzymes.org/ontology.php?ontology_id=3
Mammalian Phenotype Ontology	MP	20052305	Phenotype	https://www.ebi.ac.uk/ols/ontologies/mp
Ontology for Biomedical Investigations	OBI	27128319	Experiments	http://obi-ontology.org/

4.2.2. Résultats et discussion

4.2.2.1. Données accessibles

Les 8900 signatures toxicogénomiques actuellement déposées dans TOXsIgN sont issues de 36 publications dont 7093 proviennent du jeu de données utilisé dans le projet ChemPSy. Si les données d'Open TG-GATEs et de DrugMatrix (REF) sont d'ores et déjà intégrées dans le système, d'autres projets massifs devraient également être intégrés afin de maintenir l'attractivité de TOXsIgN. En effet, Open TG-GATEs et DrugMatrix mettent à la disposition de la communauté des données très souvent réutilisées et réorganisées dans les différentes banques du domaine (Prathipati and Mizuguchi 2016; Römer et al. 2014; Pérez, González-José, and García 2016; Mulas et al. 2017). Pour cette raison, nous travaillons actuellement à l'intégration des données de toxicogénomique issu de CMap, diXa et CEBS (Lamb et al. 2006; Hendrickx et al. 2015; Lea et al. 2017). Les données issues de ces bases permettront de diversifier l'origine de nos signatures, mais également d'avoir accès à des données de toxicogénomique non accessible via les dépôts communément utilisés comme GEO.

4.2.2.2. Soumission de signatures

En l'état actuel, TOXsIgN permet la soumission et la consultation de signatures toxicogénomiques. Cette étape de soumission est réalisée par l'intermédiaire d'un fichier Excel. L'utilisation de ce format pour l'importation des informations est pensée de façon à ce que tous les utilisateurs puissent facilement déposer un nombre important de signatures. D'autres espaces de dépôts, tels que GEO (Edgar, Domrachev, and Lash 2002), utilisent également ce principe pour la soumission de données par les utilisateurs.

Néanmoins, ce système comporte quelques limitations. L'association de la description de certains champs avec les ontologies oblige les utilisateurs à avoir accès aux termes de ces ontologies. Or, étant donné le volume de ces dernières, il est actuellement impossible de les intégrer directement dans le fichier Excel. L'utilisateur reste ainsi dépendant d'internet pour chercher les termes ontologiques dont il a besoin. Une autre limitation du fichier Excel actuel est qu'il ne permet pas de soumettre les signatures toxicogénomiques directement. En effet, celles-ci doivent être transmises ultérieurement sous la forme de trois fichiers indépendants : un pour les gènes surexprimés, un pour les gènes sous-exprimés et un dernier pour l'ensemble des gènes interrogés. Une solution actuellement envisagée est de créer un logiciel d'importation des données dans TOXsIgN sous la forme d'un tableur couplé aux ontologies utilisées pour soumettre un nouveau projet.

4.2.2.3. Confidentialité des informations

Après soumission, chaque projet est associé automatiquement à un statut « privé ». Un lien vers une publication scientifique (PubMed ID) est une condition indispensable pour faire évoluer le statut « privé » en statut « public ». Ce lien correspond au seul gage de qualité assurant que l'analyse a été évaluée par des experts du domaine. Seuls les projets « publics » sont utilisés par le système. Toutefois, TOXsIgN offre également la possibilité aux utilisateurs de soumettre des listes et de les maintenir en statut « privée » tout utilisant l'espace de travail sans qu'aucun autre utilisateur ne puisse avoir accès à ces signatures.

L'utilisation du lien vers la publication comme condition de changement de statut peut être un frein au dépôt de données. En effet, dans le cadre de l'amélioration de la reproductibilité des données en toxicologie, il est indispensable que la communauté puisse partager leurs résultats contradictoires et négatifs. Néanmoins, étant donné la difficulté de publier ce type d'informations, la plupart des signatures correspondantes ne pourront donc jamais être « publique » en raison de l'absence de publication. De plus, actuellement la méthode de transition de statut nécessite une approbation manuelle pour évaluer la validité du lien PubMed ce qui pourrait poser des problèmes à l'avenir.

4.2.2.4. Espace de travail

Grâce à l'indexation des banques Entrez gene et HomoloGene (NCBI Resource Coordinators 2016; Maglott et al. 2011), l'espace de travail de TOXsIgN permet de comparer les signatures toxicogénomiques de manière intertechnologies et interespèces. En effet, lors du dépôt des signatures génomiques, les listes d'Entrez gene IDs sont converties en listes d'identifiants HomoloGene. Ce sont ces derniers qui sont utilisés par tous les outils présents dans l'espace de travail. À ce jour, TOXsIgN n'est compatible qu'avec les listes d'Entrez gene IDs. Il conviendra également d'autoriser le dépôt de listes avec d'autres types d'identifiants, tels qu'Ensembl (Harrow et al. 2014). De plus, la conversion interespèce n'est actuellement possible que pour la vingtaine d'espèces disponible dans HomoloGene. Pour pallier à ce problème, l'intégration d'autres banques d'orthologies/homologies, telles que GeneTree (Vilella et al. 2009) et OMA (Altenhoff et al. 2015) est envisagée afin d'améliorer considérablement les comparaisons de listes appartenant à des espèces différentes – OMA couvrant à elle seule plus de 2000 espèces. Toutefois, il est important de noter que la vaste majorité de ces dernières ne constitue pas des modèles fréquemment utilisés en toxicologie.

Ces outils d'analyses et de comparaisons sont couplés à un gestionnaire de tâches permettant aux utilisateurs d'exécuter ces modules (*job*) sur la machine hébergeant le site de TOXsIgN. Une page dédiée

permet également de visualiser en temps réel l'évolution du *job* jusqu'à la production de résultats ainsi que l'historique de toutes les actions effectuées. Actuellement, ce système est parfaitement fonctionnel, mais nécessiterait de nombreuses améliorations comme, par exemple, la mise en place d'un système de file d'attente pour ne pas surcharger la machine, mais également la mise en place d'un système de répartition des ressources informatiques disponibles toujours dans un souci d'optimisation.

4.2.2.5. Exemple d'utilisation de TOXsIgN

Pour illustrer l'intérêt de l'outil TOXsIgN, j'ai utilisé l'outil de recherche et de comparaison pour démontrer la proximité de la signature toxicogénomique du fluconazole (394 mg/kg, à 6 heures d'exposition) avec d'autres composés oestrogéniques.

En utilisant l'outil de recherche avancée, j'ai cherché la signature toxicogénomique du fluconazole testé à 394 mg/kg pendant 6 heures. Cette recherche a abouti à la requête suivante (Figure 39).




Advanced search

Search parameters

AND

Select a field

Add

AND - (_type:signatures AND type:"genomic") 
AND - (_type:signatures AND tags:"fluconazole") 
AND - (_type:signatures AND _all:"394 mgkg") 

Search

Search results

Projets (No Result)

Studies (No Result)

Signatures (4)

TSS437 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to fluconazole (394 mgkg, 5 days) in the rat
(view)
Type : Genomic Project : TSP121
Organism : Rattus norvegicus Study : TSE158
Assay : TSA437

TSS441 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to fluconazole (394 mgkg, 3 days) in the rat
(view)
Type : Genomic Project : TSP121
Organism : Rattus norvegicus Study : TSE158
Assay : TSA441

TSS439 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to fluconazole (394 mgkg, .25 days) in the rat
(view)
Type : Genomic Project : TSP121
Organism : Rattus norvegicus Study : TSE158
Assay : TSA439

TSS438 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to fluconazole (394 mgkg, 1 days) in the rat
(view)
Type : Genomic Project : TSP121
Organism : Rattus norvegicus Study : TSE158
Assay : TSA438

Figure 39. Capture d'écran de l'outil de recherche avancée de TOXsIgN

Afin de récupérer la signature toxicogénomique du fluconazole à 394 mg/kg, l'outil de recherche a été utilisé. La signature de l'exposition durant 6 heures à ensuite été sélectionnée. L'identifiant de la signature est TSS439 (en rouge sur la figure).

Un bouton disponible sur la page de cette signature (TSS439) permet de directement charger celle-ci dans l'espace de travail. Cette signature, composée de 793 gènes (376 sous exprimés et 417 sur exprimés) correspondant à 751 identifiants HomoloGene, a ensuite été comparée aux autres signatures déposées dans l'espace de dépôt à l'aide de l'outil de comparaison de signature, en utilisant les paramètres par défaut (Figure 40).

Signature	r	R	n	N	Ratio	Pvalue	Zscore	Euclidean distance	Correlation distance
<input type="text"/>									
TSS205 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (10 mgkg, .25 days) in the rat	378	751	652	10770	37 %	6.82400848363682e-299	52.7519108270255	25.5929677841395	0.534376945687471
TSS440 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to fluconazole (10 mgkg, .25 days) in the rat	226	751	303	10770	27 %	6.36886760856969e-206	46.8727616291122	24.576411454889	0.471643209686015
TSS209 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to ethinylestradiol (1480 mgkg, .25 days) in the rat	395	751	907	10770	31 %	3.86865632630525e-250	45.1938705112248	29.4957624075053	0.477294182349352
TSS317 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to thioguanine (12 mgkg, .25 days) in the rat	370	751	868	10770	30 %	9.18014164343925e-227	43.0104207805062	29.9499582637439	0.446945614807702
TSS231 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to diethylstilbestrol (2.8 mgkg, .25 days) in the rat	240	751	456	10770	25 %	5.7766416913478e-165	39.1166618451114	27.2213151776324	0.397995858992561
TSS101 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to phenothiazine (386 mgkg, .25 days) in the rat	275	751	588	10770	26 %	1.0304122120797e-173	38.9653679019577	28.3725219182222	0.401742152149809
TSS325 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to raloxifene (6.5 mgkg, .25 days) in the rat	162	751	227	10770	20 %	1.6263078589976e-139	38.497966312286	25.6124969497314	0.389713731650171
TSS4913 - Open TG-GATEs - Toxicogenomics signatures of Liver after exposure to danazol (2000 mgkg, 9 hours) in the rat	216	751	391	10770	23 %	3.92362905821027e-153	38.1731674523351	26.7581763205193	0.392878608562451
TSS1181 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to beta-estradiol (150 mgkg, .25 days) in the rat	303	751	726	10770	26 %	9.46420997144378e-176	38.079982786754	29.8161030317511	0.400849743540137
TSS1070 - DrugMatrix - Toxicogenomics signatures of Liver after exposure to norethindrone (375 mgkg, .25 days) in the rat	272	751	610	10770	25 %	1.18501377064355e-164	37.5555824108014	28.8963665535998	0.388380732801847

Figure 40. Capture d'écran du résultat de comparaison (top 10) de la signature du fluconazole

Ce résultat est obtenu en comparant la signature du fluconazole 394 mg/kg, 6heures (signature sélectionnée) à toutes les signatures toxicogénomique de TOXsIgN (signature comparée). Cette page affiche les informations à propos de la signature : Identifiant et nom de la signature comparée permettant un accès direct à la page de la signature ; r : nombre d'identifiants HomoloGene en commun entre la signature sélectionnée et la signature comparée ; R, nombre d'identifiants HomoloGene présent dans la signature comparée ; n, nombre d'identifiants HomoloGene présent dans la signature sélectionnée ; N, nombre total d'HomoloGene ; Ratio, ratio r/R ; pvalue, pvalue retournée par le calcul d'enrichissement (loi hypergéométrique) ; Zscore, zscore; Euclidean distance, valeur de la distance entre la signature sélectionnée et la signature comparée basée sur les identifiants HomoloGene ; Correlation distance, valeur de la corrélation entre la signature sélectionnée et la signature comparée basée sur les identifiants HomoloGene.

La comparaison indique un enrichissement significatif ($pvalue < 1e^{-100}$) de la signature sélectionnée avec d'autres signatures de la même molécule, mais à des doses différentes (10 mg/kg pendant 6 heures). Cette analyse permet également d'identifier des similarités avec des composés oestrogéniques connus tels que l'éthinylestradiol ou le β -estradiol et d'autres ayant des propriétés oestrogéniques comme le diethylstilbestrol et le norethindrone (Lopez et al. 2013; Miyagawa, Sato, and Iguchi 2011). Ces substances sont les mêmes que celles présentes dans le groupe issu de ChemPSy, dont le fluconazole est extrait confirmant ainsi la similarité entre les signatures du fluconazole et celle de ces substances. On note également que la signature du fluconazole est très similaire à des scelles du thioguanine (traitement pour les maladies auto-immunes), du phénothiazine (antipsychotique), du raloxifene (traitement hormonal de substitution) et du danazole (molécule inhibitrice des gonadotrophines et utilisée comme médicament dans l'endométriose).

4.2.3. Conclusion et perspectives

TOXsIgN est un environnement facilitant la soumission, le stockage et la recherche en ligne de signatures toxicogénomiques grâce à l'utilisation de douze ontologies distinctes utilisées pour décrire les expériences toxicologiques et leurs résultats. L'une des caractéristiques principales de TOXsIgN est sa capacité à archiver des données hétérogènes : sans restriction d'espèces ; expériences *in vivo*, *ex vivo* ou *in vitro* ; études transgénérationnelles ; mélanges de facteurs environnementaux. Actuellement, le dépôt TOXsIgN héberge 8491 signatures toxicogénomiques provenant de 32688 expériences de puces à ADN chez l'homme, le rat, la souris et la drosophile. Chaque signature déposée est associée à un identifiant unique pour faciliter l'accès aux données et améliorer la transparence scientifique.

De nombreuses améliorations sont prévues à courts, moyens et longs termes. À court terme, la possibilité de déposer des signatures issues d'expositions à des facteurs biologiques et physiques sera développée. La mise en place de nouveaux outils pour la toxicogénomique comme l'intégration de ChemPSy ou encore d'un outil d'extraction de signatures toxicogénomiques est également envisagée dans un futur très proche. À moyen terme, la mise en place d'un système d'avertissement par mail sera mise en place afin de prévenir les utilisateurs de l'existence de listes similaires aux leurs. La refonte de l'espace de travail est également planifiée afin d'optimiser ses performances et sa facilité d'utilisation. Il pourrait ainsi être envisagé de déployer une instance Galaxy même si actuellement il est difficile de relier cet outil à une base de données autre que celles préalablement prévues. À plus long terme, de nouvelles fonctionnalités viendront enrichir l'aspect communautaire de TOXsIgN. Plusieurs utilisateurs pourront ainsi travailler sur le même projet. Nous envisageons également de mettre en place un système de notation des signatures basées sur leur fréquence d'utilisation.

TOXsIgN met à disposition des données qui intéressent fortement la communauté : les signatures toxicogénomiques. Les utilisateurs peuvent être également séduits par le potentiel de citation que représente TOXsIgN. En effet, l'outil de comparaison permet de remettre sur le « devant de la scène » des études toxicogénomiques de qualité ce qui devrait considérablement augmenter leur potentiel de citation dans les articles scientifiques. Néanmoins, pour que TOXsIgN soit reconnu par la communauté, il est nécessaire que celle-ci prenne part à la construction de ce dépôt en mettant à disposition leurs données. Pour inciter cette démarche, nous devons mettre à disposition un plus grand nombre de signatures afin que les scientifiques y trouvent un intérêt qualitatif à déposer leurs signatures. Cette démarche pourrait également être accélérée si les éditeurs encouragent vivement le dépôt de ce nouveau type de données (les signatures toxicogénomiques) au même titre que les données brutes.

Le dépôt de données dans TOXsIgN se limite actuellement aux données de toxicogénomique. Cependant, la modularité et la flexibilité de sa base n'excluent pas la possibilité de mettre en place une compatibilité avec d'autres types de données toxicologiques. Dans ce contexte, il est déjà possible de déposer des signatures physiologiques, moléculaires ou issues d'études épidémiologiques.

La problématique soulevée dans le projet TOXsIgN d'accessibilité des données analysées sous la forme de signatures nous a permis d'envisager de généraliser ce principe à l'ensemble des disciplines du vivant. Ainsi, avec mes encadrants, je coordonne également le développement d'un nouvel espace de dépôt, GeneULike, très fortement inspiré de TOXsIgN. Comme ce dernier était limité au niveau des identifiants des signatures (Entrez gene) et également sur la thématique biologique (la toxicologie), sa généralisation sous-entend détendre le format des signatures en offrant la possibilité de déposer des listes d'identifiants provenant de diverses banques (sondes nucléotidiques, protéines, gènes ou encore transcrits par exemple) et ayant trait à de nombreux sujets biologiques. Brièvement, GeneULike est un espace de dépôt multiespèces et multitechnologiques destiné à héberger des listes d'entités biologiques (gènes, transcrits, protéines, sondes nucléotidiques, ...) accessibles à la communauté scientifique via un serveur web. Ces listes sont organisées sous forme de projets associés à un identifiant unique et sont décrites par l'utilisation d'un vocabulaire contrôlé (organisme, informations, éléments méthodologiques et expérimentaux, liens PubMed vers l'article, ...). Tout comme pour TOXsIgN, un espace de travail permettra de mettre à disposition un ensemble d'outils destinés aux utilisateurs.

4.3. The ReproGenomics Viewer (RGV) : un navigateur de génome pour les données de reprogénomique

À l'heure actuelle, un des grands défis en biologie consiste à traiter, héberger et interpréter les données issues des projets utilisant les technologies de séquençage à ultra-haut débit (NGS) (Merelli et al. 2014). En effet, si les informations obtenues grâce aux puces à ADN peuvent être facilement visualisées et stockées dans des banques de données organisées autour de l'annotation des gènes, la problématique est tout autre en ce qui concerne les approches NGS. Ces dernières génèrent des données à l'échelle des bases nucléotidiques qui requièrent donc des solutions adaptées et spécifiques pour leur visualisation et leur stockage. Ce changement d'échelles est à l'origine de l'apparition de nouveaux outils tel que les « genome browsers » ou navigateurs de génome. Le plus connu d'entre eux est le navigateur de génomes généraliste de l'UCSC (Rosenbloom et al. 2014) qui héberge des données variées de génomique (transcriptomique, épigénomique, etc). Depuis quelques années et grâce à la popularisation d'outils de visualisation open source tel que JBrowse (Westesson, Skinner, and Holmes 2013), de nombreux navigateurs de génomes plus ou moins spécialisés, centrés sur une espèce donnée ou sur un type particulier de données, ont vu le jour (Choo, Heydari, et al. 2014; Heydari et al. 2013, 2014; Choo, Ang, et al. 2014; Hackenberg, Barturen, and Oliver 2011). Si de nombreuses banques hébergeant les données de puce à ADN en lien avec la reproduction ont été mises en place au cours de ces 15 dernières années (Lardenois et al. 2010; Y. Zhang et al. 2013; T.-L. Lee et al. 2009) aucun navigateur de génome dédié à cette thématique n'existait jusqu'alors. L'objectif de RGV (Darde et al. 2015) a été de mettre à disposition de la communauté scientifique d'un espace travail et de visualisation de données de NGS interspèce et dédié à la reproduction. Il se base sur l'implémentation du module JBrowse ainsi que d'un espace de travail Galaxy afin d'offrir aux utilisateurs la possibilité d'analyser des données NGS et de les visualiser via une interface intuitive.

4.3.1. Publication du ReproGenomics Viewer

The ReproGenomics Viewer: an integrative cross-species toolbox for the reproductive science community

Thomas A. Darde^{1,2}, Olivier Sallou², Emmanuelle Becker¹, Bertrand Evrard¹, Cyril Monjeaud², Yvan Le Bras², Bernard Jégou^{1,3}, Olivier Collin², Antoine D. Rolland¹ and Frédéric Chalmel^{1,*}

¹Inserm U1085-Irset, Université de Rennes 1, F-35042 Rennes, France, ²Institut de Recherche en Informatique et Systèmes Aléatoires (IRISA/INRIA) - GenOuest platform, Université de Rennes 1, F-35042 Rennes, France and

³Ecole des Hautes Études en Santé Publique, Avenue du Professeur Léon-Bernard, F-35043 Rennes, France

Received January 30, 2015; Revised March 27, 2015; Accepted April 06, 2015

ABSTRACT

We report the development of the ReproGenomics Viewer (RGV), a multi- and cross-species working environment for the visualization, mining and comparison of published omics data sets for the reproductive science community. The system currently embeds 15 published data sets related to gametogenesis from nine model organisms. Data sets have been curated and conveniently organized into broad categories including biological topics, technologies, species and publications. RGV's modular design for both organisms and genomic tools enables users to upload and compare their data with that from the data sets embedded in the system in a cross-species manner. The RGV is freely available at <http://rgv.genouest.org>.

INTRODUCTION

Sexual reproduction in eukaryotes involves a wide spectrum of biological processes by which species give rise to new individuals and thus perpetuate. These include the formation of haploid gametes after meiosis, a specific type of cell division that takes place only in the germ line. In the male, the differentiation of germ cells into highly specialized spermatozoa is a complex and tightly regulated process called spermatogenesis. This developmental process involves the sequential and coordinated expression of thousands of genes, many of them testis-specific. Spermatogenesis has thus been widely explored by several microarray-based expression studies over the last two decades (1,2) and several databases devoted to spermatogenesis and gametogenesis (3–5) or to reproduction in general (6–8) have been developed to organize and provide access to this massive quantity of data.

More recently, ultra-high-throughput next-generation sequencing (NGS) projects have imposed new challenges on the life science research community: the complex tasks of processing, hosting and interpreting these data (9). The repositories or databases referred to above, however, cannot cope with several intrinsic features of NGS data. For instance, although microarrays provide an average measurement of gene or transcript expression that can be easily displayed, NGS offers quantification at a single-base resolution, a feature that could only be observed by specific visualization tools that can take into account both genome coordinates of sequenced nucleotides and coverage information along every genomic locus. Additionally, microarray-based expression databases are typically organized around annotated entities, i.e. probes, transcripts, genes or, perhaps, corresponding proteins. Their structure is therefore incompatible with the ability of RNA-sequencing to lead to new discoveries (e.g. when new transcript isoforms are assembled and/or new loci identified) and not adapted to ChIP- or Methyl-seq analyses of specific chromatin regions, the boundaries of which cannot be strictly defined. The so-called genome browsers, a new type of database, have emerged to meet these requirements (10). UCSC's famous website (11) is a pioneer in this regard. The implementation of new modules (12,13) makes it possible to create even more flexible and intuitive browsers. These allow the hosting, visualization, customization, retrieval and analysis of various types of genomics data in a single environment, thus enabling researchers to extract and share data easily and construct new hypotheses from them. Most of these browsers, however, focus on a single species (14–17) or a single type of genomic data (18,19). To our knowledge, there is no tool directed toward a specific research field and scientific community that can bring together the major relevant studies, regardless of species and technology type.

*To whom correspondence should be addressed. Tel: +33 2 23 23 58 02; Fax: +33 2 23 23 50 55; Email: frederic.chalmel@inserm.fr

Here we present the ReproGenomics Viewer (RGV), a cross-species genomic toolbox for the reproductive community. The system is based on the implementation of a 'JBrowse genome browser' (20) and a 'Galaxy bioinformatics workflow environment' (21–23). It was developed to provide a one-stop genomic working environment and aims to assist scientists in the analysis and the mining of a wide range of high-throughput repro-genomics data, including sequencing data. RGV allows hosting, visualization and direct comparison of users' data to published genomics studies as well as to relevant genetic variations linked to reproduction. One way it does this is by enabling various genomic file format conversions. These genomic coordinates can be converted not only between genome releases of a given species but also and more importantly between different species. This key feature allows the direct comparison of data sets acquired in different organisms and thus makes RGV not only a multispecies genome browser but also a true cross-species tool for comparing reproductive genomics data. The RGV currently hosts data sets that are oriented mainly toward testis biology and spermatogenesis. In the near future, these will extend to other areas of reproduction, including gonad development, urogenital cancers and reproductive toxicology.

DESCRIPTION OF DATA SETS

As mentioned above, the RGV currently embeds 15 published studies related to male gamete development or gametogenesis in general (24–36) (Table 1). These data sets are publicly available through the NCBI Gene Expression Omnibus Repository (37). They describe the extensive re-exploration of the spermatogenesis process over the past few years by the emerging ultra-high-throughput sequencing technologies. Specifically, the studies investigated the dynamic omics landscape of developing male germ cells, including: (i) chromatin remodeling and epigenetic features such as active and repressive marks (24–25,27–30); (ii) cistromes of transcription factors important for spermatogenesis (26,29); (iii) transcriptional landscapes, defined mainly by RNA sequencing technologies (24,28,31–36); and (iv) proteomic profiles generated with the recent Proteomic Inferred by Transcriptomic approach (34). All these experiments took place in a wide spectrum of model organisms, including *Homo sapiens* (25,30,36), *Gorilla gorilla* (36), *Macaca mulatta* (36), *Mus musculus* (24–29,31–32,35), *Rattus norvegicus* (33,34), *Monodelphis domestica* (36), *Ornithorhynchus anatinus* (36), *Gallus gallus* (29,36) and *Saccharomyces cerevisiae* (as sporulation in yeast is the developmental process analogous to spermatogenesis in higher eukaryotes (38–41)). Taken together, these published data sets currently represent 342 samples, 168 of vertebrates.

In a critical step, we also gathered allele and genotype frequency data and significant genetic association findings from such public databases as GWAS and ClinVar (42,43). The control vocabulary provided by both projects enabled us to split genetic association studies into two categories: reproductive and non-reproductive symptoms. Direct links to PubMed and variant databases are provided.

THE RGV BACKBONE: DATA PROCESSING AND ORGANIZATION

The backbone of the RGV is the series of tools for processing and organizing data within the system (Figure 1A). Four types of information were manually extracted and curated for each study, including: the scientific name of each species and the genome release with which the experiments were performed and analyzed; the associated scientific publication; each biology topic investigated in the study; and the high-throughput technologies performed. Then each sample of a given data set underwent a series of automatic conversions to make it fully compatible with the RGV system (Figure 1B). Briefly, for a given sample *X* analyzed under a genome release *r-1* of Species *Y*, five processing steps were sequentially performed: (i) each of the various input data formats (bedGraph/BED, WIG, bigWig) was converted into a simple tab-delimited text file (BED); (ii) as some differences can occur even in the same genome release of a given species, the resulting BED data file might have needed to be modified to standardize, for example, the chromosome names that might differ between the Ensembl, UCSC and NCBI databases; (iii) the standardized BED file was then converted into an indexed binary format (bigWig or bw) to enable fast remote access to the data; the pairwise alignments between genome assemblies and between species provided by UCSC made it possible to convert genome coordinates in the resulting bigWig file from a genome release *r-1* of the species *Y* into (iv) the current assembly *r* of the same species *Y* and then (v) the current assembly *r* of another species *Z*.

Finally, we used manually extracted information to organize the processed data into four broad categories, i.e. biological topics, technologies, publications and species (Figure 1A). This organization is mirrored in the 'Available Tracks' option of the 'JBrowse genome browser' implemented in RGV (see the next section) to facilitate access to curated and relevant experimental data (Figure 1C).

BIOINFORMATICS TOOLS DEPLOYED IN THE RGV WORKING ENVIRONMENT

The system also integrates an implementation of the 'JBrowse genome browser' (20) and of the 'Galaxy bioinformatics workflow environment' (21–23), grafted to the RGV backbone.

The RGV working environment

To host genomics tools essential for data comparisons between genome releases and above all between species, we implemented a 'Galaxy bioinformatics workflow environment'. Briefly, Galaxy is an open web-based platform for genomic research that provides users with an easy-to-use web interface to create complex biological workflows by tools that simply need to be dragged and dropped. It is worth mentioning that the 'RGV Galaxy session' is available without creating an account. Users are, however, strongly invited to create an account to have access to their history, saved analyses, data sets and workflows. By default, this environment contains a myriad of tools designed mainly to assist

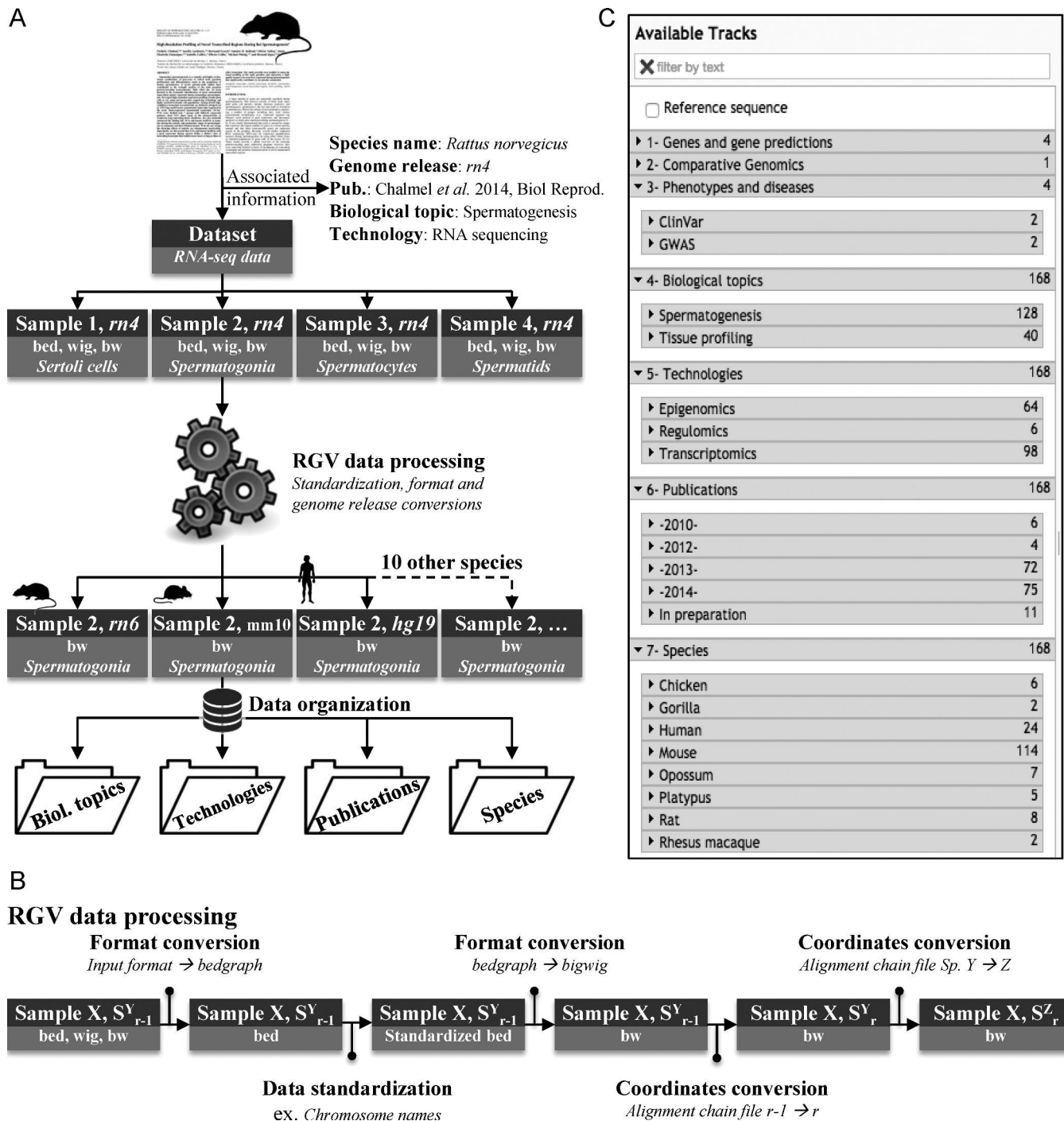


Figure 1. The RGV backbone. (A) A schematic diagram of the strategy used to process and organize each individual sample from the published data sets embedded in the RGV system. The publication by Chalmel *et al.* is taken as an example (33). The organization of the data is based on the information manually extracted from the publication (species name, genome release, biological topic and technology). (B) The 'RGV data processing' workflow used to convert data file formats, standardize data files and then to convert genome coordinates between assemblies ($r_{-1} \rightarrow r$) and between species (species $Y \rightarrow Z$). (C) Screenshot of the JBrowse 'Available tracks' menu illustrating the 'in-house' organization of the published data sets embedded in the RGV system in several categories, such as 'Biological topics', 'Technologies', 'Publications' and 'Species'.

Table 1. Published data sets relevant to gamete development currently included in the RGV system and some relevant characteristics

Publication	PubMed IDs	Species (release)	Technologies	Biological topics
Chocu <i>et al.</i> , 2014 (34)	25210130	Rat (rn4)	RNA-seq	Spermatogenesis
Hammoud <i>et al.</i> , 2014 (25)	24835570	Multi (2 species)	Chip-seq, Bisulfite-seq	Spermatogenesis
Chalmel <i>et al.</i> , 2014 (33)	24740603	Rat (rn4)	RNA-seq	Spermatogenesis
Meikar <i>et al.</i> , 2014 (35)	24554440	Mouse (mm9)	RNA-seq, smallRNA-seq	Spermatogenesis
Necsulea <i>et al.</i> , 2014 (36)	24463510	Multi (7 species)	RNA-Seq	Tissue profiling
Soumillon <i>et al.</i> , 2013 (32)	23791531	Mouse (mm9)	RNA-seq	Spermatogenesis
Erkek <i>et al.</i> , 2013 (28)	23770822	Mouse (mm9)	RNA-seq	Spermatogenesis
Gan <i>et al.</i> , 2013 (24)	23759713	Mouse (mm9)	RNA-seq, 5hMeDIP-seq	Spermatogenesis
Laiho <i>et al.</i> , 2013 (31)	23613874	Mouse (mm9)	RNA-seq	Spermatogenesis
Li <i>et al.</i> , 2013 (29)	23523368	Multi (2 species)	ChIP-seq	Spermatogenesis
Gaucher <i>et al.</i> , 2012 (26)	22922464	Mouse (mm9)	RNA-seq	Spermatogenesis
Brick <i>et al.</i> , 2012 (27)	22660327	Mouse (mm9)	ChIP-seq	Spermatogenesis
Lardenois <i>et al.</i> , 2011 (38)	21149693	Yeast (sacCer3)	Tiling Array	Sporulation (SK1, MATa-alpha)
Brykczynska <i>et al.</i> , 2010 (30)	20473313	Human (hg18)	MNase-seq	Spermatogenesis
Granovskaia <i>et al.</i> , 2010 (41)	20193063	Yeast (sacCer3)	Tiling Array	Mitosis (W101, MATa)

users in handling files; these are largely simple file manipulation tools to convert, filter, sort, select, extract features or combine files. The current release already uses this versatile Galaxy working environment to deploy two workflows.

The ‘RGV data processing’ workflow described in the RGV backbone section (Figure 1B) was conveniently implemented as a Galaxy module. This pipeline is based on the implementation of three tool suites: UCSC tools (44), bedtools (45) and CrossMap (46). The former is used for all data file format conversions in either bedGraph or bigWig formats. The second is employed for the data standardization step. Finally, the latter is used in both cross-assembly and cross-species conversions of genome coordinates and makes use of pair-wise alignment files (chain format) provided by UCSC. The entire process takes roughly 30 min for an input file (bam format) of 200 Mb. Once the conversion is completed, the user can easily upload the resulting bigWig file to the ‘RGV JBrowse session’.

A ‘genome alignment workflow’ based on the Blast-Like Alignment Tool (BLAT) (47) was implemented as a tool in the Galaxy working environment. Briefly, it allows users to use a one-step procedure to automatically align their DNA/RNA or protein sequences (fasta format) onto the 13 reference genome sequences available in the RGV system. The resulting alignments are post-processed and made available in two forms: a table including direct links to the ‘JBrowse session’ and a General Feature Format (gff) file that can be uploaded to the genome browser.

The RGV JBrowse session

As many more genomes, transcriptomes and epigenomes will be sequenced in the decade to come, a user-friendly genome browser has become essential for work in reproductive biology.

JBrowse advantages. The client-server architecture of JBrowse offers several advantages over other genome browser solutions, such as GBrowse (13): (i) the system is fully compatible with a wide spectrum of data types, including sequence files (fasta format), genomic feature files (gff), alignment files (bam) and quantitative data files (bedGraph,

wig, bigWig); (ii) genome browsing is rapid even when multiple users are processing data simultaneously; (iii) JBrowse provides a user-friendly and highly flexible graphical interface in which users can efficiently pan and zoom over a genomic sequence region and turn genomics tracks on and off by simply clicking buttons.

Track organization. As mentioned above, RGV currently includes 15 published data sets. Each has been standardized and converted via the ‘RGV data processing’ pipeline. Data were then organized into four broad categories in the JBrowse track selector, from the information manually extracted from the original publications. These categories (Figure 1C) currently include: biological topics (spermatogenesis and tissue profiling), technologies (epigenomics, regulomics or transcriptomics), publications and species (nine species).

User interaction. The implementation of JBrowse allows users to download data sets embedded into the RGV genome browser by choosing a track of interest and then by clicking on ‘Save track data’. Users can also upload their own data sets (several file formats are allowed: gff3, gtf, bigWig, bam and vcf) in the ‘JBrowse session’ to compare them to the existing tracks by using the option ‘Open’ in the ‘File’ tab. If necessary the user can first run the ‘RGV data processing’ pipeline, implemented in the ‘Galaxy session’ (see the previous section), and then upload their own tracks into the system.

Example. During spermiogenesis, sperm chromatin is remodeled into a condensed inactive state due to the replacement of histones by protamines (48,49). The latter are small arginine-rich proteins binding DNA expressed in the late-stage spermatids of many animals and plants. We used the ‘RGV JBrowse session’ to illustrate the mammalian conserved expression pattern of the genes encoding PRM1, PRM2 and PRM3 which are clustered on the human chromosome 6 (Supplementary Figure S1). Once the genes have been selected with the search bar and the genome fixed to Human (hg19), three expression data sets from human,

mouse and rat were compared (32–33,36). The corresponding tracks were accessed by (i) the ‘Publications’ tab, by selecting ‘2013>Soumillon *et al.*’, ‘2014>Necsulea *et al.*’ and ‘2014>Chalmel *et al.*’. Note that the ‘Available Tracks’ menu is organized so that the same tracks could have been identified by (ii) the ‘Technologies’ tab or by (iii) the ‘Biological topics’ tab. The examination of the displayed tracks highlights the specific post-meiotic expression of the genes encoding protamines, as well as its strong conservation across mammals.

DISCOVERING NOVEL GENES ACTIVE IN SPERMATOGENESIS

The large variety of ultra-high-throughput data across many eukaryotic organisms encourages the use of the RGV as a testing ground for building novel scientific hypotheses on the basis of relevant, curated experimental data on reproduction. The possibilities are numerous, and the applications of RGV diverse. For example, the integration and visualization of pertinent transcriptome data and genome-wide association studies related to reproductive symptoms in the ‘JBrowse session’ may help to elucidate the mechanisms through which genetic mutations lead to reproductive disorders. Another example concerns the integration of active/repressive epigenetic marks and transcriptomic data, which may help to identify the role of specific epigenetic modifications in modulating the expression of genes involved in spermatogenesis.

To corroborate RGV’s usefulness, we decided to test its ability to identify novel human loci dynamically expressed during male gamete development and conserved across species. We first integrated three RNA-sequencing studies in the ‘JBrowse session’: a tissue profiling project including samples from human testis and three other tissues (ovary, brain and placenta) published by Necsulea *et al.* (36); then we added two high-resolution expression profiles of male germ cells, one in rats (33) and the other in mice (32). Next, we analyzed the human testis sample provided by Necsulea *et al.* and assembled the transcripts with the cufflinks tool suite (50). We then sought to identify novel intergenic and multi-exonic loci that are expressed in human testes and have a meiotic and/or postmeiotic expression pattern in rodents (data not shown). This allowed us to select one promising candidate, designated TCONS_00962903, for further experimental validations to illustrate the relevance of our strategy (Figure 2A). This novel locus maps to chromosome 6 (positions 41 349 211–41 350 871) and is composed of three exons with a cumulative exon size of 659 bp. It shows preferential expression in testes compared with the other three tissue types in the study by Necsulea *et al.* (36). A simple examination of the ‘JBrowse session’, using the cross-species feature of RGV, showed very strong conservation in rodents, in which expression of this locus unambiguously peaked in spermatocytes and spermatids. This finding suggests its expression pattern in humans and rodents is similar (Figure 2A). Reverse transcriptase-polymerase chain reaction (RT-PCR) found substantial amounts of TCONS_00962903 RNA in human, mouse and rat testis samples, compared with the other tissue samples analyzed (brain, kidney, liver and lung

for rodents; epididymis, seminal vesicle and prostate for humans) and thus confirmed its ‘testis-restricted’ expression pattern (Figure 2B–D) (Supplementary file S1). Finally, as suggested by the rodent RNA-seq data, we clearly confirmed that the expression of this novel gene in the testis is restricted to the expression of the human germ cells at spermatid stage (Figure 2E).

FUTURE DEVELOPMENTS

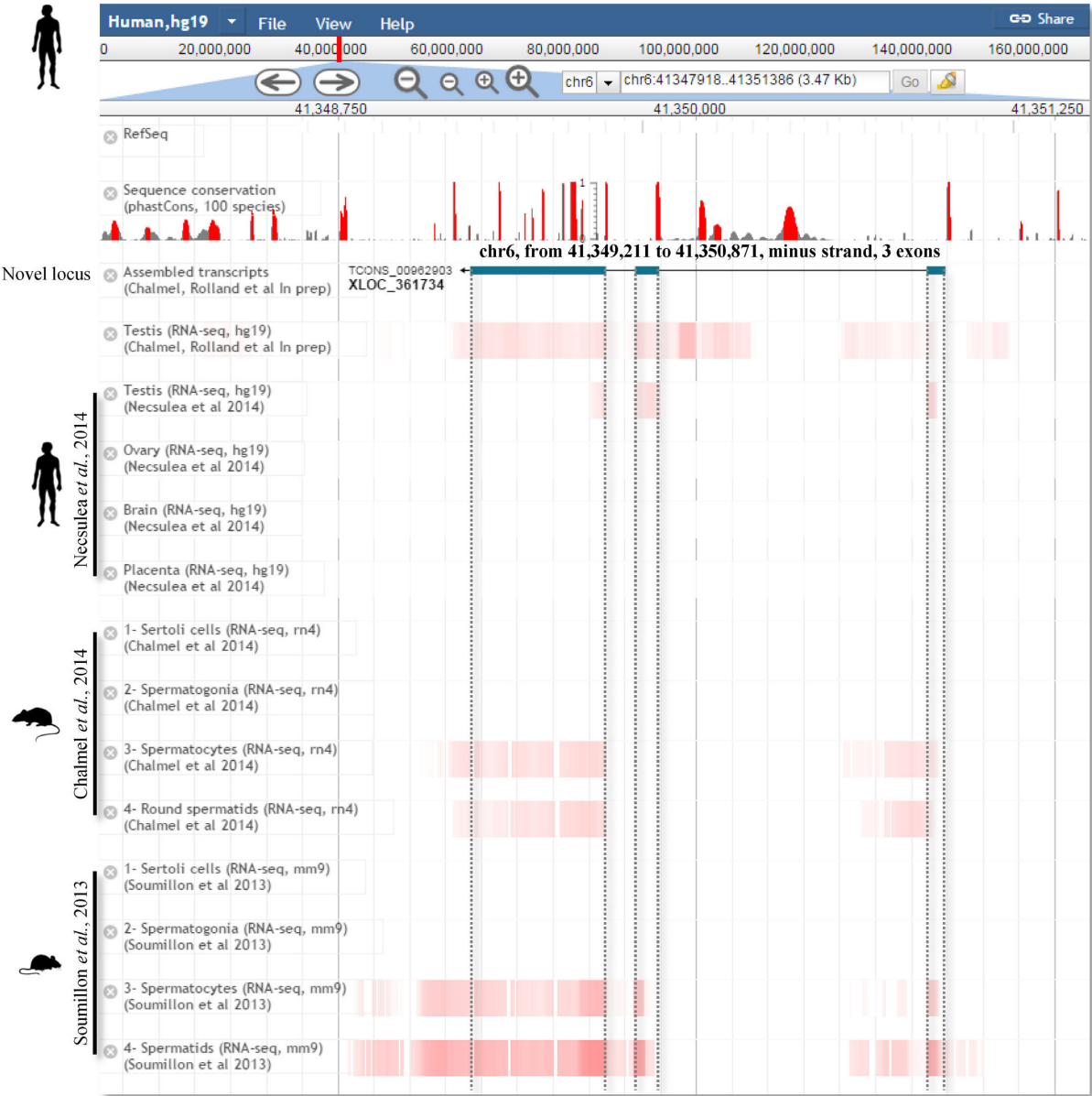
In the near future we intend to extend the scope of the RGV to keep pace with rapid technological, bioinformatics/genomic and biological/clinical advances in the reproductive sciences. In concrete terms, we are currently planning four separate actions. First, we will gather other relevant data sets from a wide range of species in RGV to cover other reproductive biological topics (e.g. gonad development, oogenesis, reproductive cancers and reproductive toxicology). We have already selected 18 studies to integrate into the system (Supplementary Table S1), and we encourage data submission from colleagues. Second, we will be adding other genetic information related to reproductive disorders (such as GWAS and Quantitative trait loci information from diverse sources and diverse model organisms). Third, we plan to develop community tools that will greatly facilitate collaborative work and stimulate the emergence of novel forms of collaboration in our research field.

Finally, we will be enhancing the features and functionalities of the ‘RGV-Galaxy working environment’. In particular, we intend to embed the ‘JBrowse genome browser’ directly into the Galaxy environment. Users will thus be able to entirely customize, and eventually share, their own personal genome browser session with their ultra-high-throughput data sets. This integration of JBrowse within the RGV-Galaxy working environment will also facilitate communications and data export between the two sessions. Another crucial point involves the direct implementation of several workflows for analysis of NGS data (e.g. RNA-seq, ChIP-seq) within the Galaxy environment. This will have several user benefits, for it will enable reproductive biologists/clinicians to perform their own analyses independently. Above all, it will help to standardize data analysis procedures within the reproductive science community to facilitate comparisons of data sets.

CONCLUSIONS

We report the development of the RGV, a webserver-based toolbox for reproductive scientists. The system combines specific solutions for ultra-high-throughput data management, curation and organization, with data conversion across releases and species (CrossMap), genome browsing (JBrowse session) and a bioinformatics workflow environment to deploy analysis pipelines (Galaxy session). RGV currently embeds 15 published data sets related to germ cell development from nine eukaryotic species. We intend to complete RGV’s repertoire with other related biological processes, other model organisms and other technologies of interest related to reproductive biology in the near future. This may help scientists and clinicians who work on reproduction to compare their own data sets to relevant

A



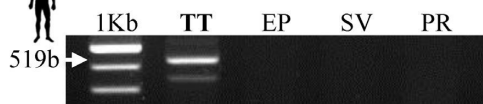
B



C



D



E



Figure 2. Tissue and cell-specific expression patterns of one novel intergenic locus are shown. (A) Structure of the novel intergenic locus (blue boxes correspond to introns), TCONS.00962903, in the human genome (release hg19), is displayed in the 'RGV JBrowse session'. Four RNA-seq data sets were selected to illustrate the transcript abundance of this promising candidate in human testes (Chalmel, F. and Rolland, A.D., in preparation) (36) as well as in rodent meiotic and post-meiotic germ cells (32,33). The amount of transcript determined in each tissue/cell and in each study is displayed as color-coded red heat maps. Red histogram bars represent the sequence conservation score distributions between 100 species as provided by the UCSC genome browser (phastCons scores, y-axis ranges from 0 to 1). TCONS.00962903 detection at the RNA level was further confirmed by RT-PCR in four rat (B) and mouse (C) tissue samples, including total testis (TT), brain (BR), kidney (KI), liver (LI) and lung (LU). RT-PCR analysis was also performed in four human tissue samples (D), including total testis (TT), epididymis (EP), seminal vesicle (SV) and prostate (PR), as well as five isolated testicular cell populations (E) including Leydig cells (LC), peritubular myoid cells (PC), Sertoli cells (SC), spermatocytes (Spc), round spermatids (rSpt) and total testis (TT) as positive control.

published studies in their specific field by overcoming the standard technical problems we face daily regarding data format, genome release and species issues. To the best of our knowledge, the RGV is the first cross-species working environment dedicated to a single biological field of interest. This community-based system could thus be applicable to other conserved biological processes studied in several model organisms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Céline Le Béguec, Yianne Ando Randriamanantena, Laetitia Guillot, François Moreews, Raphaël Charles, Aurélie Lardenois and Michael Primig for stimulating discussions and/or beta-testing RGV. We acknowledge the GenOuest bioinformatics facility for hosting the software. We also thank Dominique Mahe Poirion, Nathalie Dejucq-Rainsford and Nathalie Rioux-Leclercq for providing the human samples.

FUNDING

The Agence nationale de sécurité sanitaire de l'alimentation, de l'environnement et du travail [ANSES No. EST-13-081 to F.C.]; the Fondation pour la recherche médicale [FRM No. DBI20131228558 to F.C.]; the European Union [FEDER to F.C.]. Funding for open access charge: the Fondation pour la recherche médicale [FRM No. DBI20131228558 to F.C.].

Conflict of interest statement. None declared.

REFERENCES

- Calvel, P., Rolland, A.D., Jegou, B. and Pineau, C. (2010) Testicular postgenomics: targeting the regulation of spermatogenesis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **365**, 1481–1500.
- Rolland, A.D., Jegou, B. and Pineau, C. (2008) Testicular development and spermatogenesis: harvesting the postgenomics bounty. *Adv. Exp. Med. Biol.*, **636**, 16–41.
- Lardenois, A., Gattiker, A., Collin, O., Chalmel, F. and Primig, M. (2010) GermOnline 4.0 is a genomics gateway for germline development, meiosis and the mitotic cell cycle. *Database*, **2010**, baq030.
- Zhang, Y., Zhong, L., Xu, B., Yang, Y., Ban, R., Zhu, J., Cooke, H.J., Hao, Q. and Shi, Q. (2013) SpermatogenesisOnline 1.0: a resource for spermatogenesis based on manual literature curation and genome-wide data mining. *Nucleic Acids Res.*, **41**, D1055–D1062.
- Lee, T.L., Cheung, H.H., Claus, J., Sastry, C., Singh, S., Vu, L., Rennert, O. and Chan, W.Y. (2009) GermSAGE: a comprehensive SAGE database for transcript discovery on male germ cell development. *Nucleic Acids Res.*, **37**, D891–D897.
- Lee, T.L., Li, Y., Cheung, H.H., Claus, J., Singh, S., Sastry, C., Rennert, O.M., Lau, Y.F. and Chan, W.Y. (2010) GonadSAGE: a comprehensive SAGE database for transcript discovery on male embryonic gonad development. *Bioinformatics*, **26**, 585–586.
- Hsueh, A.J. and Rauch, R. (2012) Ovarian Kaleidoscope database: ten years and beyond. *Biol. Reprod.*, **86**, 192.
- Davies, J.A., Little, M.H., Aronow, B., Armstrong, J., Brennan, J., Lloyd-MacGilp, S., Armit, C., Harding, S., Piu, X., Roochun, Y. et al. (2012) Access and use of the GUDMAP database of genitourinary development. *Methods Mol. Biol.*, **886**, 185–201.
- Merelli, I., Perez-Sanchez, H., Gesing, S. and D'Agostino, D. (2014) High-performance computing and big data in omics-based medicine. *BioMed Res. Int.*, **2014**, 825649.
- Wang, J., Kong, L., Gao, G. and Luo, J. (2013) A brief introduction to web-based genome browsers. *Brief. Bioinform.*, **14**, 131–143.
- Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haussler, M. et al. (2015) The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.*, **43**, D670–D681.
- Goecks, J., Coraor, N., Nekrutenko, A. and Taylor, J. (2012) NGS analyses by visualization with Trackster. *Nat. Biotechnol.*, **30**, 1036–1039.
- Donlin, M.J. (2009) Using the Generic Genome Browser (GBrowse). *Curr. Protoc. Bioinform.*, Chapter 9, Unit 9.9.
- Choo, S.W., Heydari, H., Tan, T.K., Siow, C.C., Beh, C.Y., Wee, W.Y., Mutha, N.V., Wong, G.J., Ang, M.Y. and Yazdi, A.H. (2014) VibrioBase: a model for next-generation genome and annotation database development. *ScientificWorldJournal*, **2014**, 569324.
- Heydari, H., Wee, W.Y., Lokanathan, N., Hari, R., Mohamed Yusoff, A., Beh, C.Y., Yazdi, A.H., Wong, G.J., Ngeow, Y.F. and Choo, S.W. (2013) MabsBase: a Mycobacterium abscessus genome and annotation database. *PLoS one*, **8**, e62443.
- Heydari, H., Mutha, N.V., Mahmud, M.I., Siow, C.C., Wee, W.Y., Wong, G.J., Yazdi, A.H., Ang, M.Y. and Choo, S.W. (2014) StaphyloBase: a specialized genomic resource for the staphylococcal research community. *Database*, **2014**, bau010.
- Choo, S.W., Ang, M.Y., Fouladi, H., Tan, S.Y., Siow, C.C., Mutha, N.V., Heydari, H., Wee, W.Y., Vadivelu, J., Loke, M.F. et al. (2014) HelicoBase: a Helicobacter genomic resource and analysis platform. *BMC Genomics*, **15**, 600.
- Geisen, S., Barturen, G., Alganza, A.M., Hackenberg, M. and Oliver, J.L. (2014) NGSmethDB: an updated genome resource for high quality, single-cytosine resolution methylomes. *Nucleic Acids Res.*, **42**, D53–D59.
- Hackenberg, M., Barturen, G. and Oliver, J.L. (2011) NGSmethDB: a database for next-generation sequencing single-cytosine-resolution DNA methylation data. *Nucleic Acids Res.*, **39**, D75–D79.
- Skinner, M.E., Uzielov, A.V., Stein, L.D., Mungall, C.J. and Holmes, I.H. (2009) JBrowse: a next-generation genome browser. *Genome Res.*, **19**, 1630–1638.
- Blankenberg, D., Von Kuster, G., Coraor, N., Ananda, G., Lazarus, R., Mangan, M., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a web-based genome analysis tool for experimentalists. *Curr. Protoc. Mol. Biol.*, Chapter 19, Unit 19.10.1–21.
- Giardine, B., Riemer, C., Hardison, R.C., Burhans, R., Elnitski, L., Shah, P., Zhang, Y., Blankenberg, D., Albert, I., Taylor, J. et al. (2005) Galaxy: a platform for interactive large-scale genome analysis. *Genome Res.*, **15**, 1451–1455.
- Goecks, J., Nekrutenko, A. and Taylor, J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.
- Gan, H., Wen, L., Liao, S., Lin, X., Ma, T., Liu, J., Song, C.X., Wang, M., He, C., Han, C. et al. (2013) Dynamics of 5-hydroxymethylcytosine during mouse spermatogenesis. *Nat. Commun.*, **4**, 1995.
- Hammoud, S.S., Low, D.H., Yi, C., Carrell, D.T., Guccione, E. and Cairns, B.R. (2014) Chromatin and transcription transitions of mammalian adult germline stem cells and spermatogenesis. *Cell Stem Cell*, **15**, 239–253.
- Gaucher, J., Boussouar, F., Montellier, E., Curtet, S., Buchou, T., Bertrand, S., Hery, P., Jounier, S., Depaux, A., Vitte, A.L. et al. (2012) Bromodomain-dependent stage-specific male genome programming by Brdt. *EMBO J.*, **31**, 3809–3820.
- Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R.D. and Petukhova, G.V. (2012) Genetic recombination is directed away from functional genomic elements in mice. *Nature*, **485**, 642–645.
- Erkek, S., Hisano, M., Liang, C.Y., Gill, M., Murr, R., Dieker, J., Schubeler, D., van der Vlag, J., Stadler, M.B. and Peters, A.H. (2013) Molecular determinants of nucleosome retention at CpG-rich sequences in mouse spermatozoa. *Nat. Struct. Mol. Biol.*, **20**, 868–875.
- Li, X.Z., Roy, C.K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B.W., Xu, J., Moore, M.J., Schimenti, J.C., Weng, Z. et al. (2013) An ancient

- transcription factor initiates the burst of piRNA production during early meiosis in mouse testes. *Mol. Cell*, **50**, 67–81.
30. Brykczynska, U., Hisano, M., Erkek, S., Ramos, L., Oakeley, E.J., Roloff, T.C., Beisel, C., Schubeler, D., Stadler, M.B. and Peters, A.H. (2010) Repressive and active histone methylation mark distinct promoters in human and mouse spermatozoa. *Nat. Struct. Mol. Biol.*, **17**, 679–687.
31. Laiho, A., Kotaja, N., Gyenesi, A. and Sironen, A. (2013) Transcriptome profiling of the murine testis during the first wave of spermatogenesis. *PLoS one*, **8**, e61558.
32. Soumillon, M., Necseulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., Barthes, P., Kokkinaki, M., Nef, S., Gnirke, A. *et al.* (2013) Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep.*, **3**, 2179–2190.
33. Chalmel, F., Lardenois, A., Evrard, B., Rolland, A.D., Sallou, O., Dumargne, M.C., Coiffec, I., Collin, O., Primig, M. and Jegou, B. (2014) High-resolution profiling of novel transcribed regions during rat spermatogenesis. *Biol. Reprod.*, **91**, 5.
34. Chocu, S., Evrard, B., Lavigne, R., Rolland, A.D., Aubry, F., Jegou, B., Chalmel, F. and Pineau, C. (2014) Forty-four novel protein-coding loci discovered using a proteomics informed by transcriptomics (PIT) approach in rat male germ cells. *Biol. Reprod.*, **91**, 123.
35. Meikar, O., Vagin, V.V., Chalmel, F., Sostar, K., Lardenois, A., Hammell, M., Jin, Y., Da Ros, M., Wasik, K.A., Toppari, J. *et al.* (2014) An atlas of chromatoid body components. *RNA*, **20**, 483–495.
36. Necseulea, A., Soumillon, M., Warnefors, M., Liechti, A., Daish, T., Zeller, U., Baker, J.C., Grutzner, F. and Kaessmann, H. (2014) The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature*, **505**, 635–640.
37. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**, D991–D995.
38. Lardenois, A., Liu, Y., Walther, T., Chalmel, F., Evrard, B., Granovskaia, M., Chu, A., Davis, R.W., Steinmetz, L.M. and Primig, M. (2011) Execution of the meiotic noncoding RNA expression program and the onset of gametogenesis in yeast require the conserved exosome subunit Rps6. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 1058–1063.
39. Lavigne, R., Becker, E., Liu, Y., Evrard, B., Lardenois, A., Primig, M. and Pineau, C. (2012) Direct iterative protein profiling (DIPP) - an innovative method for large-scale protein detection applied to budding yeast mitosis. *Mol. Cell. Proteomics*, **11**, M111.012682.
40. Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Munster, S., Cambong, J., Guffanti, E., Stutz, F., Huber, W. and Steinmetz, L.M. (2009) Bidirectional promoters generate pervasive transcription in yeast. *Nature*, **457**, 1033–1037.
41. Granovskaia, M.V., Jensen, L.J., Ritchie, M.E., Toedling, J., Ning, Y., Bork, P., Huber, W. and Steinmetz, L.M. (2010) High-resolution transcription atlas of the mitotic cell cycle in budding yeast. *Genome Biol.*, **11**, R24.
42. Johnston, J.J., Rubinstein, W.S., Facio, F.M., Ng, D., Singh, L.N., Teer, J.K., Mullikin, J.C. and Biesecker, L.G. (2012) Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. *Am. J. Hum. Genet.*, **91**, 97–108.
43. Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorf, L. *et al.* (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**, D1001–D1006.
44. Kuhn, R.M., Haussler, D. and Kent, W.J. (2013) The UCSC genome browser and associated tools. *Brief. Bioinform.*, **14**, 144–161.
45. Quinlan, A.R. (2014) BEDTools: the Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinform.*, **47**, 11.12.11–11.12.34.
46. Zhao, H., Sun, Z., Wang, J., Huang, H., Kocher, J.P. and Wang, L. (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–1007.
47. Kent, W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.
48. Dadoune, J.P. (2003) Expression of mammalian spermatozoal nucleoproteins. *Microsc. Res. Tech.*, **61**, 56–75.
49. Balhorn, R. (2007) The protamine family of sperm nuclear proteins. *Genome Biol.*, **8**, 227.
50. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L. and Pachter, L. (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.*, **7**, 562–578.

Animal, human samples and ethical considerations

Adult male Sprague-Dawley rats and male C57Bl/6 mice were used for tissue collection (testis, brain, kidney, liver and lung) and RT-PCR experiments. They were purchased from Elevage Janvier (Le Genest-Saint-Isle, France). Animal experiments were performed in conformity with the principles for the use and care of laboratory animals and in compliance with French and European regulations on animal welfare.

Human materials were obtained at Rennes University Hospital from patients: normal testis and epididymis samples were collected at autopsy; normal seminal vesicles were collected from patients who underwent radical prostatectomy; prostate tissues were obtained from healthy men who underwent prostatic adenomectomy for benign prostatic hyperplasia. The local Ethics Committee approved the study protocol, "Study of normal and pathological human spermatogenesis", registered under n° PFS09-015 at the French Biomedicine Agency. Human tissues used for RT-PCR were frozen in liquid nitrogen and stored at -80°C until analysis. Human testicular cells (spermatids, spermatocytes, Sertoli cells, peritubular myoid cells, and Leydig cells) were isolated from human testes as described in (Chalmel *et al.*, 2007) (Chalmel, Rolland *et al.*, in prep.).

RNA isolation and RT-PCR

RNA was isolated using the RNeasy miniextraction kit (Qiagen) and treated with DNase (Promega). Complementary DNA was obtained from 500 ng of total RNA using the High capacity RNA-to-cDNA Kit (Applied Biosystem). Conventional PCR was performed using *Taq* polymerase (Qiagen) and a Peltier thermocycler (Labgene). Forward and reverse primers are listed below (overlined in yellow) for human, mice and rat TCONS respectively. Sequences and product sizes (overlined in blue) are indicated.

Human

GGGA CAGGAGCCCTAGGCAATATG TATTAAGTAGACTTTTCATGAACTCAACGAGATACTGCAAGAGATAAAATAA
TGAGATGCTACGTTCATAGGATCTGTCTCTGAGTCCACATTGTGGGAGAAAAGACCATTCTTCGGAAAGATGCAGA
GGAAGCCCCGTACCTCAAGGGATTGAGATGGGAGACTGAGGGAATCCTGGGGTCTCTGCCCTCTCCCTCACTGGG
AAACAAGCTCTGTTTCTTCATGTGCTGGAGCCCTGCCTCTGTGGGTGCAGCCTGAGGACAAAGGCAGTGTGCAT
CCCCTGATCCCCTGCCCCCTACAACCTCCTGCCAGCCTCTTGGGTGCTCCTCAGCTTGTCGTCAGGGAGGAAGC
AGCCATCTCCTCCGATGACTGACTTCAAGCCTCAACTCTCCCACAGCTTCTCAAAGCCAAGGCTGAGGCCAAGT
CCTAGCTGCTGCCTTGGGACAGGTCAAGGGCCTCCTGCAATCTCATGCACAAA CTGGTACCCATGAGCATGTGGC
CTAAGAGATTTATTGACCCCTCTCCAAGTCCTTGCCAGAGGAATCCTGCACACCTATATCCACAATGAAGGGAGC
CATATCCCCAGCTGGGTAGCCATA

Forward primer: CAGGAGCCCTAGGCAATATG

Reverse primer: CACATGCTCATGGGTACCAG

SEQUENCE SIZE: 624

PRODUCT SIZE: 519

Mouse

GACATCACAGCAGGGA GAGGGGCCCTTCACAACCTTT CCCACTTGGACTTTCAAGGACTCAGAAAGCTAGTGCTA
GAGAAAGAAGATCATGTCCTTGGGAAAGATGGAGAATGAGTTCGAAGTCAGGAAAGCCCCGACTCTACAAGGGACT
GAGATGGAAGACCCAGACCCAGGGCGACCCGCACTCCCACTCACCTGGGAATGAGCTCTGCCCCCTCCACGGAGCA
AGCTTCAAAGGTCTGCCTGTGTGGCTGCAGCTTGCGGACAAAAGCGTCTCCACCCCAGGCCGGATCCAGTCCC
TCTTCTAGTCTCTCGGGGGCCGTCCAGCTTGCCATCAGAGAGGAAGACACTGGTTTCCAAGGAGACAGTCTTTCC
AGCTCAGCTCTCCCACTCTCCGAAGGAAAAAGCTGAGGGCAAGTCCCAGCTGTTATCTTGGCGGGATCCGGT
ACCTCTGCGCCAGCTGGTGCCTATGAGCGTGTGGCCCAAGGGCTTTATTTCATCCACTCCCCAAACCTTTGCCAGG

GCAAGCTCTGAGCACCTTTATCCATGTTGAAGGGGTTACACATCCCTAGGGCAGTATCTCCGGTGACATCACTGGT
GGAAGATAACAAGTCTTCTATGCCACCTTTACTAAAATGAACGTGCCTGCTCAATTAATTGCTATTTATCTGTGT
CTACTGTAAACACGCAAACTGTCCATTTTTAACTGAGAATATATTTTTATCCTCAATGTGAACCCAGTTGGGAC
TTCCAGAATACTGTATCCCAGTGTATCTTTCAAATATATTTTTATTTTATTTGTGCATGT

Forward primer: GAGGGGCCCTTCACAACCTTT
Reverse primer: CACACGCTCATAGGCACCAG

SEQUENCE SIZE: 812
PRODUCT SIZE: 467

Rat

GACATCACAGAAGGG GAGGGGCCCTTCACAACCTTG TCCCGCTTGGACTTTCAAGGACTCAGAAAGCTAGTGCTAG
AGAAAGAAGAGCATGTCCTTGGGAAAAATGGAGGATGATTTCGAAGTCAGGAAAGCCCGACTTTACAAGGGGTTG
AGGTGGAAGACCCAGACCCAGGACGTCCAGCACTCCCACTCACCTGGGAATGAGCTCTGCCCCCTCCACGGAACAC
GCTTCAAAGGTCTGCCTGTGTGGGTGCAGCTTGCGGACAAAAGTGGTCCTCCATCCCAGGCCGGACCCACTTCCG
GCCCCCTCGGGTGCCTCCCAGCTTACCATCGGAGAGGAAGACACTGGTTTTCCAGGGATACAGTTTTTCCCGTTCAG
CTCTCGAATGACTCATTGAAGGAAAAAGCTGAGGGTAAGCCCCAGCTGTTACCTTGGCGTGACCCGGTACCTCTG
CGCCAG CTGGTGCCCATGAGCGTGTG GCCCAAGGGCTTTATTGACCCGCTCCCCAAGCTTTTGCCAGAGAAAGCT
TTGAGCACCTTATCCATATTGAAGGGATTACATCCCCAGGGCAGTATCCCCCGTGATGGCACTGGGGAAAGAT
ACCAAATCCTCTCTGCCACCTTTACTGAAATGAGCGTGTCCACTCAATTAATCACTAATTTTCTGTATCTACTGT
AAACATATGAAGCTATTTTAATGGAGAATATATTTATATTCTCATTATGACCCCAAGTGGGGACTTCAGAACACT
GTATCCCAATGTATCTTTCAAATACATTT

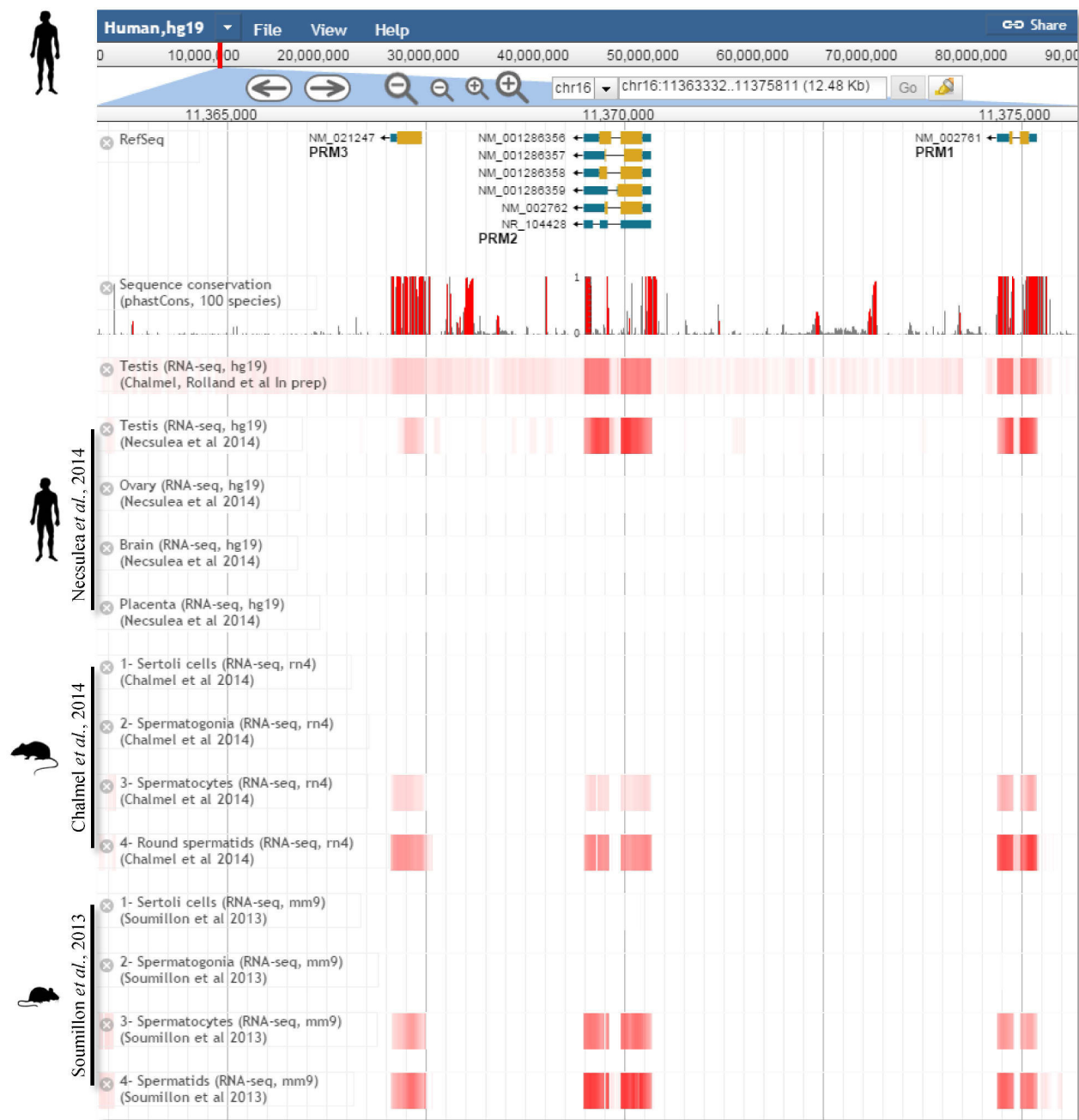
Forward primer: GAGGGGCCCTTCACAACCTTG
Reverse primer: CACACGCTCATGGGCACCAG

SEQUENCE SIZE: 780
PRODUCT SIZE: 461

Supplementary Table S1 - Published datasets relevant for the reproductive science community selected for integration within the RGV system.

Publications	Pubmed IDs	Species (release)
Zimmermann et al.	Submitted	Mouse (mm9)
Chalmel, Rolland et al.	In prep.	Human (hg19)
Becker et al.	In prep.	Yeast, sacCer3
Mu et al., 2014	25228648	Mouse (mm9)
Wang et al., 2014	24813617	Mouse (mm10)
Margolin et al., 2014	24438502	Mouse (mm9)
Smagulova et al., 2013	23870400	Mouse (mm9)
Jiang et al., 2013	23663777	Zebrafish (danRer7)
Yokobayashi et al., 2013	23486062	Mouse (mm9)
Ng et al., 2013	23352811	Mouse (mm8)
Shen et al., 2012	23235881	Mouse (mm9)
Sleutels et al., 2012	22709888	Mouse (mm9)
Khil et al., 2012	22367190	Mouse (mm9)
Lavigne et al., 2012	21997732	Yeast (sacCer3)
Tan et al., 2011	21925322	Mouse (mm9)
Smagulova et al., 2011	21460839	Mouse (mm9)
Bernstein et al., 2010	20944595	Human (hg19)
Xu et al., 2009	19169243	Yeast, sacCer3

Supplementary Figure S1 - Tissue and cell-specific expression patterns of protamines. Structure of the genes encoding protamines (PRM1-3) (blue and orange boxes correspond to untranslated and coding exons respectively) in the human genome (release hg19), is displayed in the *RGV JBrowse session*. Four RNA-seq datasets were selected to illustrate the transcript abundance of these genes in human testes (Necsulea *et al.*, 2014; Chalmel, Rolland *et al.*, in preparation) as well as in rodent meiotic and post-meiotic germ cells (Soumillon *et al.*, 2013; Chalmel *et al.*, 2014).



4.3.2. Résultats

4.3.2.1. Version antérieure de RGV

RGV est un projet initié avant mon arrivée au sein du laboratoire. Lorsque j'ai débuté ma thèse, ce navigateur de génome dédié à la reproduction reposait sur l'outil GBrowse (Generic Genome Browser) (Stein et al. 2002) et indexait 5 études dans 4 espèces différentes (Homme, rat, souris et levure). GBrowse est un outil de visualisation d'annotations de génomes développé dans le cadre du projet GMOD (Generic Model Organism Database). Il embarque les mêmes fonctionnalités que Jbrowse (outils de navigation sur le génome, visualisation de données quantitatives et qualitatives, importation de fichiers, ...), ce dernier étant lui-même un dérivé de GBrowse. Toutefois, GBrowse repose sur une technologie plus difficile à maintenir dans le temps et surtout plus coûteuse en ressources informatiques, ce qui entraînait d'importants ralentissements lors de la navigation. Par ailleurs JBrowse présente un affichage plus clair et plus agréable d'un point de vue graphique. J'ai ainsi procédé à la migration de RGV de GBrowse vers Jbrowse.

4.3.2.2. Données stockées

Depuis sa publication, de nombreuses études accessibles au public via le dépôt GEO ont été ajoutées à RGV. Au total le site indexe plus de trente études traitant de la spermatogenèse et du développement des gamètes en général par des approches de séquençage à ultra-haut débit (NGS) (Tableau XVII). Plus spécifiquement, ces études portent sur: (i) la dynamique transcriptionnelle au cours du développement des cellules germinales mâles ; (ii) le remodelage de la chromatine et des marques épigénétiques dans ces mêmes cellules ; (iii) le cistrome de facteurs de transcription importants pour la spermatogenèse; et (iv) des approches de protéogénomique couplant reconstruction de transcrits par RNA-seq et identification des produits de gènes correspondants par spectrométrie de masse de type MS/MS. Enfin, j'ai également intégré (v) les données génotypiques issues des bases de données publiques GWAS (Welter et al. 2014) et ClinVar (NCBI Resource Coordinators 2015). Toutes ces expériences ont été réalisées chez un large éventail d'organismes modèles, dont l'Homme, le gorille, le singe, la souris, le rat, l'opossum, l'ornithorynque, la poule et la levure. Pris ensemble, ces données représentent 507 échantillons, dont plus de 240 chez les vertébrés.

Publication	PMID	Espèce	Technologie	Sujet biologique
Chocu et al., 2014	25210130	Rat	RNA-seq	Spermatogenesis
Hammoud et al., 2014	24835570	Multi	Chip-seq, Bisulfite-seq	Spermatogenesis
Chalmel et al., 2014	24740603	Rat	RNA-seq	Spermatogenesis
Meikar et al., 2014	24554440	Souris	RNA-seq, smallRNA-seq	Spermatogenesis
Necsulea et al., 2014	24463510	Multi	RNA-Seq	Tissue profiling
Soumillon et al., 2013	23791531	Souris	RNA-seq	Spermatogenesis
Erkek et al., 2013	23770822	Souris	RNA-seq	Spermatogenesis
Gan et al., 2013	23759713	Souris	RNA-seq, 5hMeDIP-seq	Spermatogenesis
Laiho et al., 2013	23613874	Souris	RNA-seq	Spermatogenesis
Li et al., 2013	23523368	Multi	ChIP-seq	Spermatogenesis
Gaucher et al., 2012	22922464	Souris	RNA-seq	Spermatogenesis
Brick et al., 2012	22660327	Souris	ChIP-seq	Spermatogenesis
Tan et al., 2011	21925322	Souris	ChIP-seq	Spermatogenesis
Sleutels et al., 2012	22709888	Souris	ChIP-seq	Spermatogenesis
Ng et al., 2013	23352811	Souris	ChIP-seq	Tissue profiling
Yokobayashi et al., 2013	23486062	Souris	RNA-seq	Spermatogenesis
Margolin et al., 2014	24438502	Souris	RNA-seq	Tissue profiling
Howarth et al., 2014	24874946	Souris	RNA-seq	Spermatogenesis
Korfanty et al., 2014	25450459	Souris	ChIP-seq	Spermatogenesis
Soh et al., 2015	26378784	Souris	RNA-seq	Tissue profiling
Smagulova et al., 2011	21460839	Souris	ChIP-seq	Tissue profiling
Lardenois et al., 2011	21149693	Levure	Tiling Array	Sporulation
Brykczynska et al., 2010	20473313	Homme	MNase-seq	Spermatogenesis
Granovskaia et al., 2010	20193063	Levure	Tiling Array	Mitosis
Zimmermann et al., 2015	25710594	Souris	RNA-seq	Spermatogenesis
Khil et al., 2012	22367190	Souris	ChIP-seq	Spermatogenesis
Xu et al., 2009	19169243	Levure	Tiling Array	Sporulation
Lavigne et al., 2012	21997732	Levure	DIPP	Sporulation

Tableau XVII. Liste des études disponibles dans RGV

4.3.2.3. Outils de conversion

Le cœur de RGV repose sur l'enchaînement spécifique d'outils pour le traitement et l'organisation des données au sein du système. Tout d'abord, quatre types d'informations doivent être extraites et organisées manuellement pour chaque étude : le nom de l'espèce et la version du génome à partir de laquelle les analyses ont été réalisées ; la publication scientifique associée ; la thématique biologique de l'étude ; les technologies à haut débit utilisée (Figure 41A). Pour chaque échantillon une série de conversions automatiques a ensuite été mise en place afin de rendre les données correspondantes compatibles avec le système RGV (Figure 41B). Ainsi, pour un échantillon X d'une espèce Y et une version de génome r-1, cinq étapes de traitement sont successivement réalisées: (i) Les données d'entrée (bedGraph / bed, wig, bigWig) sont converties en simple fichier texte tabulé (bed); (ii) le fichier bed est normalisé afin, par exemple, d'homogénéiser les noms de chromosomes qui peuvent être différents entre les bases de données Ensembl, UCSC et NCBI ; (iii) le fichier bed standardisé est ensuite converti en un format de fichier binaire indexé (bigWig ou bw) pour permettre un accès rapide aux données; Enfin, les fichiers d'alignements entre les différentes versions d'un même génome et entre les génomes d'espèces différentes (chain files fournis par UCSC) sont utilisés afin de convertir les coordonnées de la version r-1 du génome de l'espèce Y dans (iv) la version actuelle r du génome de la même espèce Y, puis (v) dans la version actuelle r du génome d'une espèce Z. Les informations extraites ensuite manuellement organisées en quatre grandes catégories (les sujets biologiques, les technologies, les publications et les espèces) afin de permettre aux utilisateurs d'explorer les données et de sélectionner celles qu'ils souhaitent afficher.

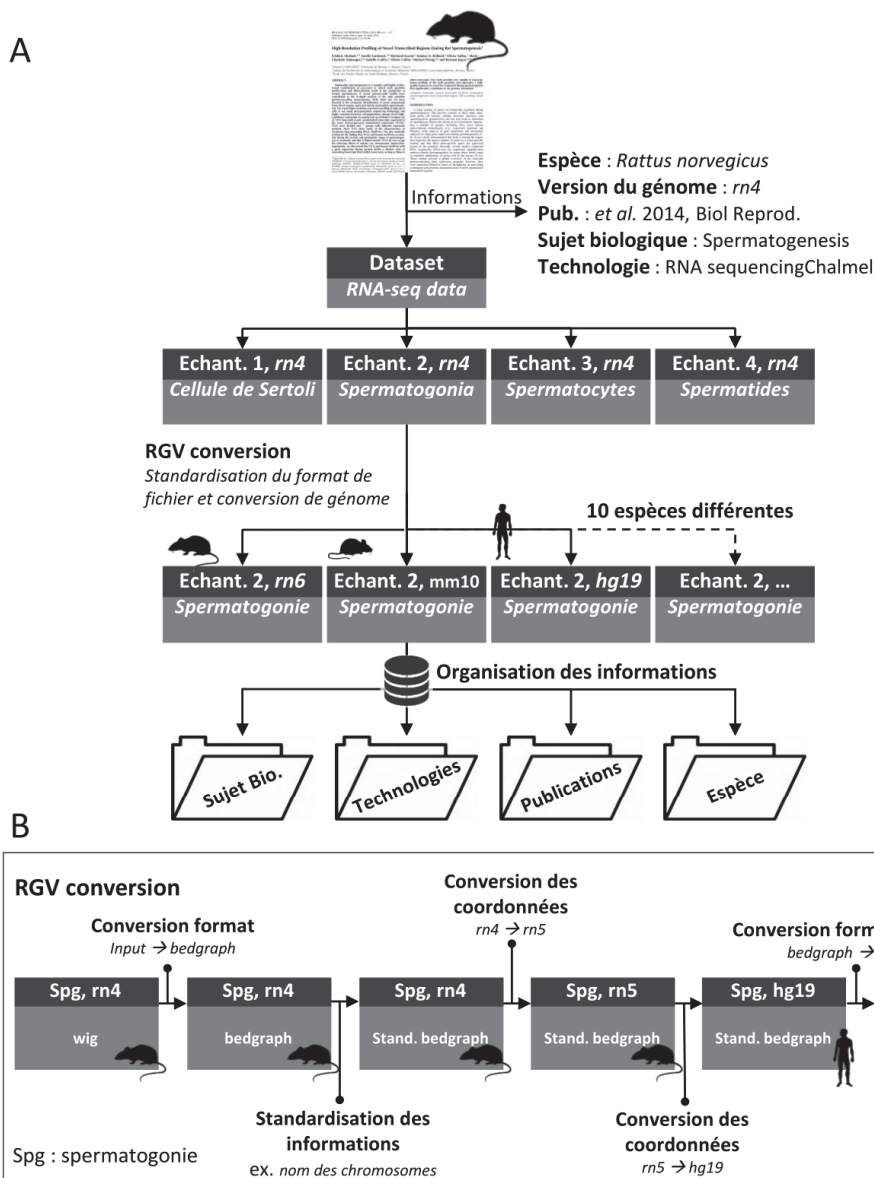


Figure 41. Outil de conversion de RGV

Processus d'indexation des informations dans RGV. A- Organisation des informations. Pour un set de données téléchargé, les informations associées (auteur, publication, espèce, version de génome, ...) sont extraites et classées selon quatre niveaux (publication, espèce, technologie et sujet biologique). Avant d'être stocké sur RGV, chaque échantillon du jeu de données subit un processus de conversion. B – Conversion des données. Les données d'entrée sont converties en simple fichier bedgraph puis normalisées (homogénéisation des informations). Les fichiers d'alignements entre les différentes versions d'un même génome et entre les génomes d'espèces différentes (chain files fournis par UCSC) sont utilisés afin de convertir les coordonnées de la version r-1 (rn4) du génome de l'espèce Y (rat) dans la version actuelle r (rn5) du génome de la même espèce Y (rat), puis dans la version actuelle r (hg19) du génome d'une espèce Z (Homme).

4.3.3. Discussion et perspectives

The ReproGenomics Viewer est navigateur de génomes spécialisé autour d'une thématique biologique, la reproduction. Il présente cependant une caractéristique unique par rapport aux autres outils du genre en permettant de comparer des études menées chez des espèces différentes et de visualiser les résultats sur l'un des 13 génomes indexés.

À l'heure actuelle, je travaille sur la mise en place d'un certain nombre d'améliorations permettant de faire évoluer RGV, aussi bien en ce qui concerne les données intégrées, la manière dont elles sont traitées ou encore leur visualisation :

Dans cette nouvelle version, la variété des sujets indexés sera étendue à l'oogenèse, à la différenciation des gonades, aux cancers de la sphère urogénitale ainsi qu'à la reprotoxicologie. Pour faire face au nombre croissant d'études et à l'augmentation exponentielle des conversions qui devront être réalisées, ces dernières seront limitées aux espèces modèles telles que l'Homme, le rat, la souris, le chien, la vache et le poisson-zèbre.

Actuellement la conversion d'une espèce à une autre est dépendante de l'existence des fichiers *chain* présent sur le site de l'UCSC. Le passage d'un génome à un autre peut nécessiter la conversion vers un ou deux génomes intermédiaires si le fichier *chain* n'existe pas. Par exemple la conversion d'échantillon obtenu sur le génome de la vache (bosTau) version bosTau7 vers la version rn6 du rat nécessite une conversion au préalable de bosTau7 vers rn4 puis de rn4 vers rn5 et enfin de rn5 vers rn6. En plus d'être chronophage, ces conversions accumulent les pertes d'informations au fur et à mesure des enchainements. Ceci est d'autant plus vrais quand on passe par une espèce intermédiaire. Pour pallier à ce problème, je travaille avec l'aide de l'équipe bioinformatique de l'UCSC, sur la mise en place de processus automatisé pour la création des fichiers *chain* quand ceux-ci n'existe pas.

Dans sa version précédente, RGV était par ailleurs limité aux études pour lesquelles les auteurs fournissaient des fichiers bed. Dans la nouvelle version ce sont les fichiers de données brutes de séquences qui seront utilisés : cela permettra d'inclure plus d'études (non plus limitées à celles fournissant des fichiers bed) tout en permettant d'uniformiser le traitement des données. L'harmonisation du protocole d'analyse permettra de plus de proposer pour chaque nouvelle piste du navigateur de génome un affichage brin spécifique, indispensable notamment pour discriminer les données d'expression de gènes chevauchants localisés sur des brins différents de l'ADN. De plus, une visualisation des profils d'expression à l'échelle du gène ou du transcrit sera rendue possible. Toujours en termes de visualisation, nous travaillons actuellement à rendre possible l'affichage de données de type cellules uniques (Single-Cell RNA-seq).

Enfin, un dernier point d'amélioration de RGV vise à corriger la dissociation entre l'espace de travail (Galaxy) et le navigateur de génome (Jbrowse). Actuellement il n'est en effet pas possible de visualiser directement dans Jbrowse les fichiers générés sous Galaxy : il faut préalablement télécharger le fichier de résultat puis le charger dans le navigateur. L'intégration de RGV directement au sein du JBrowse de Galaxy, comme cela est par exemple le cas pour le navigateur Trackster (Goecks et al. 2012), permettrait de résoudre ce problème. Le navigateur de génome RGV s'inscrit dans le cadre de ma thèse puisqu'il constitue un outil potentiellement intégrable à l'espace de dépôt TOXsIgN. Bien que n'ayant pas été initialement conçu comme tel, l'utilité de cet outil appliqué à la toxicogénomique sera indéniable au vu de la démocratisation des approches de NGS dans ce domaine.

5. Autre publication



Original article

PepPSy: a web server to prioritize gene products in experimental and biocuration workflows

Olivier Sallou¹, Paula D. Duek², Thomas A. Darde^{1,3}, Olivier Collin¹,
Lydie Lane^{2,4,†} and Frédéric Chalmel^{3,†,*}

¹Genouest Bioinformatics Platform, IRISA, Campus de Beaulieu, Rennes 35042, France, ²CALIPHO Group, SIB Swiss Institute of Bioinformatics, CMU, Michel Servet 1, Geneva 1211, Switzerland, ³IRSET, Inserm U1085, 9 avenue du Professeur Léon Bernard, Rennes 35000, France and ⁴Department of Human Protein Sciences, Faculty of Medicine, University of Geneva, CMU, Michel Servet 1, Geneva 1211, Switzerland

*Corresponding author: Tel: +33 02 2323 5802; Fax: +33 02 2323 5055; Email: frederic.chalmel@inserm.fr

†These authors contributed equally to this work.

Received 26 November 2015; Revised 17 March 2016; Accepted 13 April 2016

Abstract

Among the 20 000 human gene products predicted from genome annotation, about 3000 still lack validation at protein level. We developed PepPSy, a user-friendly gene expression-based prioritization system, to help investigators to determine in which human tissues they should look for an unseen protein. PepPSy can also be used by biocurators to revisit the annotation of specific categories of proteins based on the 'omics' data housed by the system. In this study, it was used to prioritize 21 dubious protein-coding genes among the 616 annotated in neXtProt for reannotation. PepPSy is freely available at <http://peppsy.genouest.org>.

Database URL: <http://peppsy.genouest.org>.

Introduction

Analysis of the human genome led to the identification of approximately 20 000 protein-coding genes, which produce a variety of functional proteoforms via different mechanisms including genetic polymorphisms, alternative splicing, post-translational modifications, or processing. The Human Proteome Project (HPP) launched by the Human Proteome Organization (HUPO) aims at providing experimental validation for these proteoforms and understanding their role in health and disease (1). neXtProt is an innovative knowledge

platform focusing on human proteins, that is built on top of UniProtKB/Swiss-Prot annotations (2) and provides additional expert-curated information on protein expression, subcellular localization, post-translational modifications and protein variations, gathered from selected high-throughput datasets (3). Whenever possible, neXtProt provides links to primary data providers, such as the Human Protein Atlas (HPA) (4). It is also cross-referenced to the main expert-curated sequence knowledgebases UniProt (2) and RefSeq (5), and to integrative databases such as GeneCards (6) that

automatically collect available information on human genes. Since neXtProt has been chosen as the reference database for the HPP, it collects and displays all the mass spectrometry and antibody-based data generated by the HPP consortium (7).

Since 2008, UniProtKB has been assigning a ‘protein existence’ (PE) score to each entry, ranging from PE1 (entry with evidence at protein level) to PE5 (uncertain). PE5 entries correspond to dubious protein sequences that are kept in UniProtKB/Swiss-Prot until other major genome annotation resources involved in the CCDS project (8) reach a consensus and annotate them either as coding or as noncoding. If the sequence is annotated as noncoding, the corresponding entry is removed from UniProtKB and stored in UniParc (9). If it is annotated as coding, then it can take a PE score of 1 (existence validated at protein level), 2 (existence validated at transcript level), 3 (existence validated by homology), or 4 (existence validated by a gene model), depending on the available information. neXtProt uses the same classification and upgrades PE2, PE3 and PE4 entries to PE1 when there is clear mass spectrometry evidence for the existence of the corresponding gene products. According to the standard metrics table of the neXtProt database (release 2014-09-19), there are still about 3000 genes that were classified as coding but for which none of the products could be validated at protein level (3). Most of these proteins escaped detection either because their physico-chemical properties are not compatible with mass-spectrometry or antibody-based techniques, or because their expression pattern is restricted in time or space. One of the first objectives of the Chromosome-centric Human Proteome Project (C-HPP) is to validate the existence of those so-called ‘missing’ proteins corresponding to neXtProt entries annotated with PE scores ranging from PE2 to PE4 (7).

Customizable workflows, such as CAPER 2.0 (10), have already been used to discover novel genes by integrating transcriptomic and proteomic data (11). In the HPP context, the use of transcriptomics datasets has also proven to be valuable to determine where to look for missing proteins (12–14). To the best of our knowledge, there is no dedicated and freely-accessible workflow to assist researchers in prioritizing missing proteins for detection, and suggesting potential biological samples. We thus developed PepPSy, a user-friendly gene expression-based system aiming at identifying the tissues in which unseen gene products should be looked for. PepPSy currently embeds eight filtration criteria and seven prioritization modules dedicated to the filtration of protein candidates and their ranking according to gene expression-based information.

Since most PE5 entries correspond to erroneous translations of non-protein-coding elements, they are not expected to be detected in proteomics experiments. Therefore, any claim for a PE5 protein identification is a

priori considered as artefactual and needs to be carefully documented (15). However, recent publications highlighted that a few of them might still be true proteins (16,17). In order to get a chance to be validated by proteomics, it is crucial that these entries are kept in UniProtKB, the safest way being by upgrading their PE status to PE2–4. We have applied PepPSy to revisit the annotation of PE5 entries based on existing expression information. As a result, 21 entries have been provided transcriptomic evidence for further validation and reclassification by neXtProt and UniProt, and eight entries have been reclassified.

PepPSy data and interface

The ‘Filtration’ tab

PepPSy has a simple and intuitive interface including a ‘Filtration’ tab (Figure 1A) which enables users to apply up to eight selection criteria from menus organized into four categories: *General information*, *Transcriptomic and proteomic data*, *Annotation data* and *Search your own proteins*. In the former category, the user can select candidates according to: (i) the neXtProt database release; (ii) the chromosome location provided by neXtProt; (iii) the neXtProt protein existence (PE) status; and, (iv) the observability status (ranging from ‘observable’ to ‘unobservable’) according to the Farrah *et al.* classification that is supposed to reflect the probability to identify a given protein based on several parameters (18). In the *Transcriptomic and proteomic data* category, PepPSy allows users to filter candidate gene products based on (v) their presence in six distinct transcriptomic and proteomic datasets including: the NCBI UniGene/EST dataset (45 human experimental conditions (HECs), such as tissues, organs or cell types) (5); an Affymetrix 3’ Gene array dataset (109 HECs; Supplementary Table S1) (19); an Affymetrix All Exon array dataset containing 12 human tissues (raw data available on manufacturer’s website at http://www.affymetrix.com/support/technical/sample_data/exon_array_data.affx); the Illumina RNA-sequencing dataset (32 HECs) of the Human Protein Atlas (HPA) database (4); the antibody-based protein expression profiling (83 HECs) of HPA (4); and a LC/MS-based protein expression profiling (30 HECs) (20). One can also directly select candidates expressed in (vi) specific anatomical systems or tissues. In the corresponding menu, HECs that are analysed in at least one of the 6 transcriptomic and proteomic datasets are classified using the human anatomy ontology provided by the neXtProt platform (ftp://ftp.nextprot.org/pub/current_release/controlled_vocabularies/caloha.obo). Briefly, the menu contains 2 levels in which 60 human tissues/organs (level 2) are classified in 17 major systems of the human body (level 1).

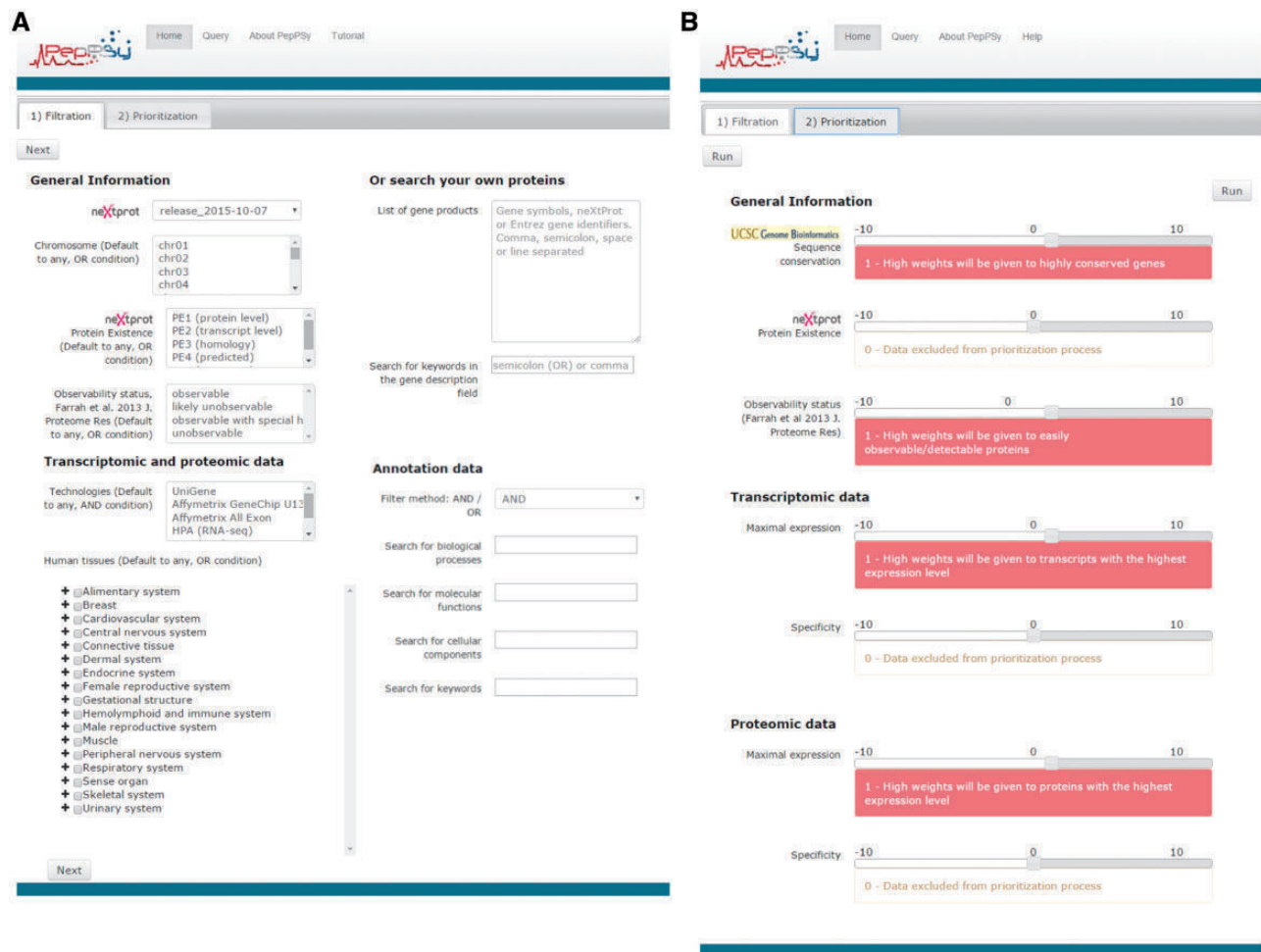


Figure 1. PepPsy layouts. (A) The 'Filtration' and (B) the 'Prioritization' tab contain menus to filter proteins and sliders to increase (red, from +1 to +10) or decrease (blue, from -10 to -1) contribution of each prioritization module to the final ranking.

The *Annotation data* category allows users to select candidates according to (vii) their associated annotations including biological process, molecular function and cellular component GO terms as well as keywords that are searched in the 'Keyword' field of each entry (21) as provided by neXtProt (3). Finally, (viii) a custom list of gene products can be queried using the text field of the latter category termed 'Search your own proteins'. The User is also able to select candidates by searching keywords in the 'description' field of each entry. Currently, PepPsy only accepts neXtProt and NCBI's Entrez Gene identifiers as well as gene symbols. For other database entry conversions, users are referred to the UniProt's online conversion service available at <http://www.uniprot.org/uploadlists> (9).

The 'Prioritization' tab

PepPsy provides a direct access to seven different ranking modules in a 'Prioritization' tab organized into three

categories (Figure 1B), i.e. *General Information*, *Transcriptomic data* and *Proteomic data*.

The *General Information* category contains three modules which allow to obtain a ranked list of the neXtProt entries according to: (i) their degree of evolutionary sequence conservation across vertebrate genomes by averaging the base-by-base phastCons conservation scores calculated among 100 vertebrate species as provided by the UCSC genome browser (22); (ii) their neXtProt PE status; and, (iii) their observability status as defined by Farrah and collaborators (18).

Both *Transcriptomic data* and *Proteomic data* categories include two modules, *Maximal expression* and *Specificity*, which rank neXtProt gene products according to their maximal abundance at the (iv) transcript- and (v) protein-levels by using the transcriptomic and proteomic datasets described in the 'Filtration' tab; and to their restricted expression pattern across human tissues at the (vi) transcript- and (vii) protein-levels as defined by the Shannon entropy Q (categorical) (19,23).

Importantly, the interface of the *Prioritization* tab enables users to apply higher or lower weights to each prioritization module thus increasing or decreasing its contribution to the final ranking algorithm. By taking the weight applied to the *Specificity* module of the *Transcriptomics data* category as an example: a high positive weight (from +1 to +10) will tend to give a better ranking to gene products displaying an expression at the transcript-level (based on the UniGene/EST, the Affymetrix 3' Gene array, the Affymetrix All Exon array and/or the Illumina RNA-sequencing transcriptomic datasets) restricted to only few human tissues than to gene products ubiquitously detected in all organs; a low negative weight (from -10 to -1) will give a high final rank to ubiquitously-expressed genes; finally, a weight of 0 means that this module will be not taken into account to compute the overall ranking. Another example: a high positive weight (from +1 to +10) to the *Sequence conservation* module will tend to give high importance in the overall ranking to neXtProt entries associated with protein sequence conserved across 100 vertebrate species.

By default, the **output page** displays the top 20 neXtProt entries but users can change this setting as they deem appropriate. The result is displayed in the form of a table (Figure 2A) containing one protein entry per line with columns for: neXtProt IDs (hyperlinked to the neXtProt knowledgebase); gene symbols (hyperlinked to the HGNC database (24)) and protein descriptions; NCBI Entrez gene IDs (hyperlinked to the NCBI website); the color-coded evolution of the neXtProt PE status over time; the current PE status; the observability status according to the classification published by Farrah *et al.* (18); the rank of each neXtProt entry computed by the PepPSy prioritization system based on the weight scheme defined by the user; and the human tissues in which the corresponding gene products display the highest abundance based on the six distinct transcriptomic and proteomic datasets (NCBI-UniGene, Affymetrix 3' array and All Exon, Illumina RNA sequencing, HPA antibody-based and HPM LC/MS-based protein expression profiles) (Figure 2A). Results can be exported as an archive file containing both a complete (full) and a lighter (light) text files (tabular format) reporting the entire protein list and corresponding ranking information via a 'click here' link at the top of the page and subsequently imported into an Excel sheet.

In addition to the output table, a graphical interactive interface (Figure 2B), called Body map, is conveniently available by clicking on the stickman to the left of each neXtProt entry (Figure 2A). This tool may help users to summarize and explore the current knowledge regarding the expression at the transcript- and protein-level of each individual gene product in all major tissues of the human

body through the six transcriptomic and proteomic technologies hosted in the system. Briefly, a menu on the right side contained 3 levels in which HECs (level 3, blue and italic font style) are organized in human tissues/organs (level 2) classified in major systems of the human body (level 1). Human systems, tissues and HECs are color-coded according to the expression levels of the considered gene product, measured by the six transcriptomic and proteomic technologies and displayed as first, second and third quartile values (Figure 2B) (cf. Design and Implementation). The expression values of a gene product at the system level (level 1) are inferred from the maximum expression values observed in the tissues (level 2) belonging to the same system which themselves are inferred from the maximum expression values measured in the HECs (level 3) belonging to the same tissues.

Finally, it is worth mentioning that the welcome page includes a link to a brief tutorial for PepPSy.

Design and implementation

Transcriptomic and proteomic data pre-processing

For each neXtProt entry in each human experimental condition (HEC), the number of ESTs from the UniGene dataset (5), intensity values from the Affymetrix 3' Gene array and All Exon datasets (25), FPKM values from the RNA-seq dataset provided by HPA (4) and protein expression levels provided by Kim and coworkers in the HPM database (20) were extracted from the different datasets, log2-transformed and quantile-normalized to facilitate comparison between HECs. Staining intensity information provided by HPA were associated with staining intensity values of 0, 1, 2 or 3 for not-detected, low, medium or high staining intensities, respectively.

Finally, the resulting transcriptomic and proteomic expression values were discretized into four levels using quartile values: low (< first quartile, Q1), medium (< Q2), high (< Q3) and very high (> Q3).

Weight scheme and overall prioritization

The prioritization parameters enable users defining their own weight combination or scheme reflecting the contributions of each module to the overall prioritization. This highly supple feature allows for complex queries pertaining to very specific biological questions.

As already described in the gene prioritization system (GPSy) (19), the integration of transcriptomic and proteomic datasets with distinct ranking strategies forms the basis of PepPSy's modular architecture allowing for

maximum query flexibility. In PepPSy, the precomputation of module-wise ranks greatly accelerates the process of prioritization. To combine the ranking output of each individual module, the absolute rank of each neXtProt entry in a ranked list (module i , protein j) is transformed into relative ranks using the formula:

$$rR_{i,j} = \frac{r_{i,j}}{R_i}$$

where $r_{i,j}$ and $rR_{i,j}$ are the absolute and relative ranks of the protein respectively, and R_i is the total number of entries in the ranked list.

When the system is queried, candidate protein entries in the input list are mapped onto the pre-computed ranked lists. An overall rank $r\bar{R}_j$ of a given neXtProt identifier j is computed based on an inter-module weighted average of the individual module ranks:

$$r\bar{R}_j = \frac{\sum_{i=1}^I w_i \times rR_{i,j}}{\sum_{i=1}^I w_i}$$

where w_i is the weight applied to each module selected in the user interface. The final output is a reordered list based on the overall ranking of each gene product entry.

Technical issues and updates

Above the PepPSy database generated by Python and Tcl/Tk scripts, there is a web application implemented in PHP with the Symfony framework. The application indexes the input files using the Lucene library. This index allows to search in the whole elements with multiple input cross parameters. When user selects information, the application searches in the index for matching gene product identifiers and calls the Tcl/Tk scripts with the selected prioritization parameters and the matching proteins. Results (JSON format) are then returned and displayed as a table.

As PepPSy is closely related to the neXtProt knowledgebase, information for each gene product entry is updated and processed every time a novel neXtProt release is made available to the community. Note that PepPSy offers the possibility to query the system and to prioritize lists of gene products based on the information of the older neXtProt releases.

Application to PE5 protein reannotation

The PepPSy interface was used to search for solid transcript expression information relative to the 616 entries labelled as ‘dubious’ (PE5) in the neXtProt release

2014-09-19. In the Filtration ‘tab’, we selected 2014-09-19 as release date, PE5 as neXtProt protein evidence, and the four transcriptomics datasets. Default parameters were used for the ‘Prioritization’ step. The result was ten entries for which there was information in all the four transcriptomics datasets. Table 1 shows the prioritized list of these ten entries, with the corresponding gene symbols, observability status, and the two tissues in which they have the highest expression level according to each dataset.

The first entry on this list (NX_B1AH88) corresponds to a very unusual annotation case in UniProtKB/Swiss-Prot (hence neXtProt). Although the usual UniProtKB/Swiss-Prot procedure is to merge all the splice isoforms that arise from one gene into a single UniProtKB/Swiss-Prot entry, this particular isoform has been annotated as a separate entry because it results from another reading frame and does not share any sequence with the other isoform (NX_P30536) (26). Because NX_B1AH88 and NX_P30536 are potentially transcribed from the same gene, they are mapped to the same Ensembl gene identifier (ENSG00000100300) and inherit any transcriptomics data linked to this gene identifier. Therefore, the data that was retrieved for the dubious NX_B1AH88 isoform is probably an artefact and would need to be remapped to the well-known isoform (NX_P30536, PE1).

The second entry from the list, NX_Q9NPU4, corresponds to the C14orf132 gene. There is a complete consensus across the four transcriptomics datasets showing that C14orf132 is expressed at highest levels in the brain. Initially, the annotation resources had predicted that this gene would encode a 173 aa protein. After reexamination, they chose another reading frame, resulting in a 83 aa transmembrane protein, that would be conserved in most mammalian species. The sequence has been changed in UniProtKB (04-MAR-2015 release) and neXtProt (2015-04-28 release). Given the available transcriptomics data, one should look for it in brain samples, if possible after membrane enrichment. Since trypsin cleavage would lead to a single, hydrophobic peptide, a special methodology may be required for its detection. Until a conclusive proof for its existence at protein level is established, the status of the entry has been changed to PE3 (validated by homology).

For two other entries, NX_Q8N5Q1 (FAM71E2, line 9) and NX_Q9BWV7 (TTLL2, line 4), there is also a clear consensus among the four transcriptomics datasets, indicating highest expression levels in testis. Because cDNAs for FAM71E2 have been found in different tissues (thymus, testis and brain), the status of the entry has been changed to PE2 (validated at transcript level) in UniProtKB (04-MAR-2015 release). Since two unique peptides corresponding to FAM71E2 were identified by mass spectrometry in sperm (27), neXtProt reclassified the corresponding

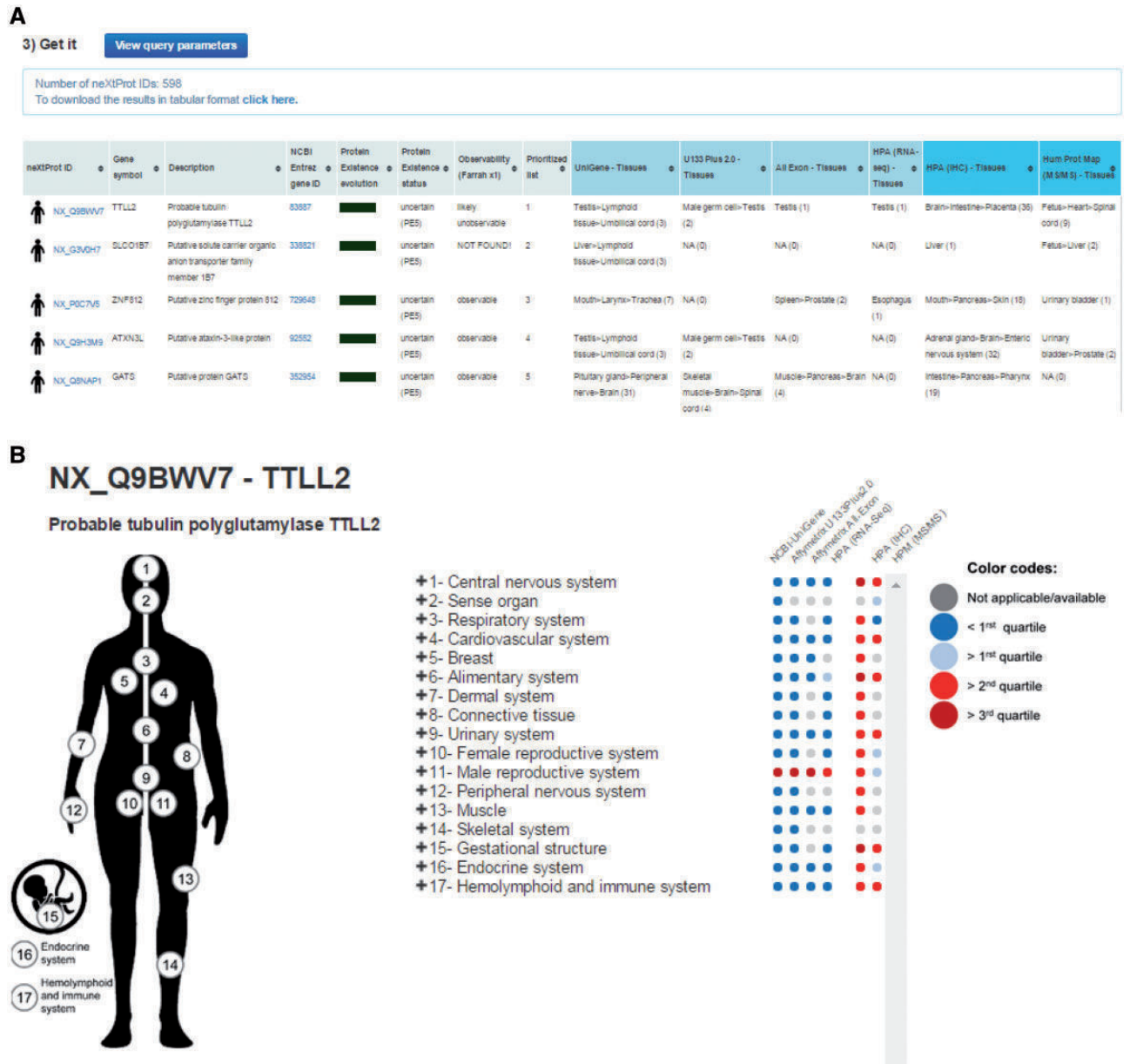


Figure 2. PepPSy outputs. (A) The output displays columns for neXtProt IDs, gene symbols, protein description, NCBI Entrez gene IDs, a color-coded evolution of the neXtProt Protein Existence status over time, the current Protein Existence status, the *observability* status, the prioritized rank of each gene product and the human tissues in which ranked genes are expressed in the different transcriptomic and proteomic datasets.

entry as PE1. The TTLL2 gene has a mouse ortholog, cDNAs have been found in testis, and the CCDS consortium has recently reclassified it as protein-coding. Therefore, NX_Q9BWV7 is currently under examination by UniProtKB/Swiss-Prot and neXtProt curators for upgrading to PE2 (validated at transcript level). The peptide GGLDAPDCLPYDSLSFTSR, which uniquely maps on the corresponding NX_Q9BWV7 entry has been identified in testis by mass spectrometry. Since a single peptide is not sufficient to upgrade a protein to PE1 (validated at protein level), targeted LC-SRM studies on testis, using other peptides will need to be performed.

For NX_Q96SF2 (CCT8L2, line 6), three datasets out of four show highest expression levels in testis. The CCT8L2 gene, only found in Human and Chimp, is thought to have arisen by duplication in the Hominoidea lineage after its divergence from the Cercopithecidae (28). Although some mass spectrometry information is available (20), it has not passed the stringent criteria quality of the HPP for validation. Until we gain a more conclusive proof for its existence at protein level, the status of the entry has been changed to PE2 (validated at transcript level). Given the available transcriptomics data, it would be wise to look for this protein in testis-related samples. However, its

Table 1. List of the ten PE5 entries from neXtProt release 2014-09-19 that have expression information in the four transcriptomic datasets

neXtProt accession	Gene symbol	Observability status	UniGene	Affymetrix U133 Plus 2.0	Affymetrix All Exon 1.0	Hum. Protein Atlas RNA-seq
1 NX_B1AH88	TSPO	observable	Intestine > Brain > ...	Mouth > Bone marrow > ...	Spleen > Intestine > ...	Bone marrow > Skin > ...
2 NX_Q9NPU4	C14orf132	with special handling	Brain > Testis > ...	Brain > Spinal cord > ...	Brain	Brain > Oviduct > ...
3 NX_P0CF97	FAM200B	likely unobservable	Brain > Eye > ...	Male germ cell > Brain > ...	Brain	Ovary > Kidney > ...
4 NX_Q9BWW7	TTL2	likely unobservable	Testis > ...	Male germ cell > Testis	Testis	Testis
5 NX_Q5T036	FAM120AOS	likely unobservable	Brain > Lung > ...	Placenta > Pituitary gland > ...	Intestine > Kidney > ...	Placenta > Thyroid > ...
6 NX_Q96SF2	CCT8L2	likely unobservable	Testis > Brain > ...	Male germ cell > Testis	Intestine > Pancreas > ...	Testis
7 NX_Q8IVY1	C1orf210	with special handling	Pancreas > Intestine > ...	Female germ cell	Intestine > Kidney > ...	Intestine > Stomach > ...
8 NX_P0CB46	CASP16	likely unobservable	Spleen > Uterus > ...	Female germ cell	Intestine	Intestine
9 NX_Q8N5Q1	FAM71E2	observable	Testis > Thymus > ...	Male germ cell	Testis	Testis
10 NX_Q96HZ7	C21orf119	likely unobservable	Intestine > Prostate > ...	Testis > Male germ cell > ...	Testis > Kidney > ...	Skeletal muscle > Thyroid > ...

The list has been prioritized using the default parameters of PepSy. The four last columns show the two tissues in which the highest expression levels have been reported in each dataset.

definitive validation by mass spectrometry may not be an easy task since it differs from its closest paralog CCT8L1P by only a few residues. This is probably why its observability status has been set as ‘likely unobservable’ by Farrah *et al.* (18).

For NX_Q8IVY1 (C1orf210, line 7), two of the datasets indicate highest expression levels in intestine. C1orf210 has a clear ortholog in mouse (2610528J11Rik, UniProtKB Q9CQM1) and has been identified by mass spectrometry in fetal liver, pancreas, prostate (29), breast (30) and ovary (31). It has been shown to be phosphorylated on Tyr-94 in human cell lines of various origins (32). Therefore, its status has been changed to PE1 (validated at protein level).

For NX_P0CF97 (FAM200B, line 3), the datasets indicate expression in various tissues, including brain. FAM200B is well conserved among mammals, and the CCDS consortium has recently classified it as protein coding. Although some mass spectrometry information is available (20), it has not passed the stringent criteria quality of the HPP. Until a more conclusive proof for its existence at protein level is available, its status has been changed to PE3 (validated by homology). However, given the high sequence similarity with FAM200A, a conclusive proof for its existence at protein level will be hard to find.

For NX_Q5T036 (FAM120AOS, line 5), two of the datasets indicate expression in placenta. It is classified as protein coding by the CCDS consortium, is conserved in several mammalian species, but has not been detected by mass spectrometry yet. Given the available transcriptomics data, it would be wise to look for this protein in placenta. Until a conclusive proof for its existence at protein level is available, this entry is currently under examination by UniProtKB/Swiss-Prot curators for upgrading to PE3 (validated by homology) or 4 (predicted).

For NX_Q96HZ7 (C21orf119, line 10), two datasets show expression in testis. However, C21orf119 is predicted to encode a long non-coding RNA. Therefore, NX_Q96HZ7 will remain PE5 until contradictory evidence is available.

For NX_P0CB46 (CASP16, line 8), two datasets show expression in intestine. However, a consensus has been reached between the different resources to classify it as a pseudogene (not protein coding). Therefore, NX_P0CB46 has been deleted from UniProtKB/Swiss-Prot and neXtProt.

In conclusion, the use of PepSy as a biocuration companion tool has allowed to quickly prioritize 10 PE5 proteins for re-annotation by biocurators, among a total of 616 proteins. As shown in Table 2, two of them have been validated at protein level (C1orf210 and FAM71E2), and five have been or will be upgraded to PE1-3. For four of

Table 2. Summary of the reannotation of PE5 entries

neXtProt accession	Gene symbol	New PE	Sample suggestion for further analyses
NX_B1AH88	TSPO	PE5	
NX_Q9NPU4	C14orf132	PE3	Brain; membrane fraction
NX_P0CF97	FAM200B	PE3	Brain; will be difficult to distinguish from FAM200A
NX_Q9BWV7	TTLL2	PE1? (in progress)	Testis
NX_Q5T036	FAM120AOS	PE3? (in progress)	Placenta?
NX_Q96SF2	CCT8L2	PE2	Testis; will be difficult to distinguish from CCT8L1P
NX_Q8IVY1	C1orf210	PE1	
NX_P0CB46	CASP16	Deleted	
NX_Q8N5Q1	FAM71E2	PE1	
NX_Q96HZ7	C21orf119	PE5	
NX_Q8IYS8	BOD1L2	PE2	Testis
NX_P0CG32	ZCCHC18	PE3	Brain?; will be difficult to distinguish from ZCCHC12
NX_C9J798	RASA4B	PE3	Skeletal muscle; quasi undistinguishable from RASA4

them, further unambiguous proof of their existence needs to be found by mass spectrometry or Ab-based proteomics. Table 2 indicates in which sample these proteins could be investigated.

We will continue to provide UniProtKB/Swiss-Prot curators with transcriptomic evidence and other available information for all PE5 entries with transcriptional information in at least one of the available transcriptomics datasets (433 entries in total). We have started the process with the 11 PE5 entries which have RNA-seq information in HPA and EST information in UniGene, as well as information in one of the two microarray datasets. Following this work, the status of three entries (NX_Q8IYS8, BOD1L2; NX_P0CG32, ZCCHC18 and NX_C9J798, RASA4B) has been upgraded to PE2-3. According to the three datasets, BOD1L2 highest levels are found in testis. Interestingly, BOD1L2 was unambiguously identified by two peptides detected by mass spectrometry in spermatozoa samples (27). Therefore, the entry will probably be upgraded to PE1 soon. For ZCCHC18 and RASA4B, there is no clear consensus between the transcriptomic datasets, and even with the right sample, it will be difficult to unambiguously validate their existence at protein level due to strong similarity with ZCCHC12 (NX_Q6PEW1) and RASA4 (NX_O43374), respectively.

Conclusion

PepPSy has been developed as a user-friendly gene expression-based prioritization system, to help investigators to determine in which human tissues they should look for an unseen protein and curators to quickly look at available transcriptomics data for a list of protein. In this work, PepPSy has been applied to prioritize twenty-one proteins annotated as ‘Uncertain’ (PE5) in UniProtKB/Swiss-Prot and neXtProt for revision. As a result, 21 proteins have

been provided transcriptomic evidence and biocurators have changed the status of eight of these based on all available information. PepPSy can now be used to choose the samples in which to look for the seven proteins that have been reclassified as PE2 or PE3, and to identify potentially problematic cases. In the near future, PepPSy will be used to revise the annotation of the 412 remaining PE5 entries with transcriptional information in at least one of the available transcriptomics datasets. Therefore, PepPSy has been revealed as an efficient companion tool for neXtProt biocuration and quality management workflows, and for the C-HPP project which aims to get an accurate picture of the validation status of all human protein coding genes. In the near future we will extend the scope of PepPSy to stay abreast of rapid technological advances. We will gather other relevant datasets in PepPSy to cover other biological topics by including other tissues, specific cell types from single-cell RNA-seq data, and chemical-induced/disease-associated experimental samples. Finally, we are also currently planning to develop a community tool embedded in PepPSy that will stimulate the annotation of other missing proteins by facilitating collaborative work.

Supplementary data

Supplementary data are available at Database Online.

Acknowledgements

The authors thank Antoine D. Rolland, Ramona Britto, Emmanuelle Becker, Laëtitiya Guillot and the GenOuest bioinformatics facility for stimulating discussion as well as for beta-testing this web server. They also thank Lionel Breuza and Sylvain Poux from the UniProt group at SIB for their input and their work in updating proteins in UniProtKB/Swiss-Prot. More generally, they thank all the curators from UniProt and the CCDS consortia for their dedication in providing up-to-date high-quality annotations for human genes and proteins, thus providing neXtProt with a solid foundation.

Funding

This work was supported by the ‘Agence nationale de sécurité sanitaire de l’alimentation, de l’environnement et du travail’ (ANSES) [grant number EST-13-081] and the ‘Fondation pour la recherche médicale’ (FRM) [grant number DBI20131228558] awarded to F.C. This project also benefited from European Union financial help (FEDER) [grant number 14MF434-01]. neXtProt development benefits from extensive funding support from the SIB Swiss Institute of Bioinformatics. The neXtProt server is hosted by VitalIT, the bioinformatics competence center that supports and collaborates with life scientists in Switzerland. Funding for open access charge: Institut national de la santé et de la recherche médicale (Inserm).

Conflict of interest. None declared.

References

1. Legrain, P., Aebersold, R., Archakov, A. *et al.* (2011) The human proteome project: current state and future direction. *Mol. Cell Proteomics*, 10, M111 009993.
2. Breuza, L., Poux, S., Estreicher, A. *et al.* (2016) The UniProtKB guide to the human proteome. *Database (Oxford)*, 2016,
3. Gaudet, P., Michel, P.A., Zahn-Zabal, M. *et al.* (2015) The neXtProt knowledgebase on human proteins: current status. *Nucleic Acids Res.*, 43, D764–D770.
4. Uhlen, M., Fagerberg, L., Hallstrom, B.M. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, 347, 1260419.
5. NCBI Resource, C. (2016) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, 44, D7–D19.
6. Safran, M., Dalah, I., Alexander, J. *et al.* (2010) GeneCards Version 3: the human gene integrator. *Database (Oxford)*, 2010, baq020.
7. Omenn, G.S., Lane, L., Lundberg, E.K. *et al.* (2015) Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *J. Proteome Res.*, 14, 3452–3460.
8. Farrell, C.M., O’leary, N.A., Harte, R.A. *et al.* (2014) Current status and new features of the Consensus Coding Sequence database. *Nucleic Acids Res.*, 42, D865–D872.
9. Pundir, S., Magrane, M., Martin, M.J. *et al.* (2015) Searching and navigating UniProt databases. *Curr. Protoc. Bioinf.*, 50, 1 27 21–21 27 10.
10. Wang, D., Liu, Z., Guo, F. *et al.* (2014) CAPER 2.0: an interactive, configurable, and extensible workflow-based platform to analyze data sets from the Chromosome-centric Human Proteome Project. *J. Proteome Res.*, 13, 99–106.
11. Wu, P., Zhang, H., Lin, W. *et al.* (2014) Discovery of novel genes and gene isoforms by integrating transcriptomic and proteomic profiling from mouse liver. *J. Proteome Res.*, 13, 2409–2419.
12. Chalmel, F. and Rolland, A.D. (2015) Linking transcriptomics and proteomics in spermatogenesis. *Reproduction*, 150, R149–R157.
13. Diez, P., Droste, C., Degano, R.M. *et al.* (2015) Integration of Proteomics and Transcriptomics Data Sets for the Analysis of a Lymphoma B-Cell Line in the Context of the Chromosome-Centric Human Proteome Project. *J. Proteome Res.*, 14, 3530–3540.
14. Segura, V., Medina-Aunon, J.A., Mora, M.I. *et al.* (2014) Surfing transcriptomic landscapes. A step beyond the annotation of chromosome 16 proteome. *J. Proteome Res.*, 13, 158–172.
15. Bruford, E.A., Lane, L. and Harrow, J. (2015) Devising a consensus framework for validation of novel human coding loci. *J. Proteome Res.*, 14(12), 4945–8.
16. Dong, Q., Menon, R., Omenn, G.S. *et al.* (2015) Structural bioinformatics inspection of neXtProt PE5 proteins in the human proteome. *J. Proteome Res.*, 14, 3750–3761.
17. Carapito, C., Lane, L., Benama, M. *et al.* (2015) Computational and mass-spectrometry-based workflow for the discovery and validation of missing human proteins: application to chromosomes 2 and 14. *J. Proteome Res.*, 14, 3621–3634.
18. Farrah, T., Deutsch, E.W., Hoopmann, M.R. *et al.* (2013) The state of the human proteome in 2012 as viewed through PeptideAtlas. *J. Proteome Res.*, 12, 162–171.
19. Britto, R., Sallou, O., Collin, O. *et al.* (2012) GPSy: a cross-species gene prioritization system for conserved biological processes—application in male gamete development. *Nucleic Acids Res.*, 40, W458–W465.
20. Kim, M.S., Pinto, S.M., Getnet, D. *et al.* (2014) A draft map of the human proteome. *Nature*, 509, 575–581.
21. Huntley, R.P., Sawford, T., Mutowo-Meullenet, P. *et al.* (2014) The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res.*, 43(Database issue):D1057–63.
22. Karolchik, D., Barber, G.P., Casper, J. *et al.* (2014) The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res.*, 42, D764–D770.
23. Schug, J., Schuller, W.P., Kappen, C. *et al.* (2005) Promoter features related to tissue specificity as measured by Shannon entropy. *Genome Biol.*, 6, R33.
24. Gray, K.A., Seal, R.L., Tweedie, S. *et al.* (2016) A review of the new HGNC gene family resource. *Hum. Genomics*, 10, 6.
25. Chalmel, F., Lardinois, A., Evrard, B. *et al.* (2012) Global human tissue profiling and protein network analysis reveals distinct levels of transcriptional germline-specificity and identifies target genes for male infertility. *Hum. Reprod.*, 27, 3233–3248.
26. Lin, D., Chang, Y.J., Strauss, J.F. 3rd. *et al.* (1993) The human peripheral benzodiazepine receptor gene: cloning and characterization of alternative splicing in normal tissues and in a patient with congenital lipoid adrenal hyperplasia. *Genomics*, 18, 643–650.
27. Jumeau, F., Com, E., Lane, L. *et al.* (2015) Human spermatozoa as a model for detecting missing proteins in the context of the chromosome-centric human proteome project. *J. Proteome Res.*, 14, 3606–3620.
28. Mukherjee, K., Conway de Macario, E., Macario, A.J. *et al.* (2010) Chaperonin genes on the rise: new divergent classes and intense duplication in human and other vertebrate genomes. *BMC Evol. Biol.*, 10, 64.
29. Pinto, S.M., Manda, S.S., Kim, M.S. *et al.* (2014) Functional annotation of proteome encoded by human chromosome 22. *J. Proteome Res.*, 13, 2749–2760.
30. Edwards, N.J., Oberti, M., Thangudu, R.R. *et al.* (2015) The CPTAC data portal: a resource for cancer proteomics research. *J. Proteome Res.*, 14, 2707–2713.

-
31. Mertins,P., Yang,F., Liu,T. *et al.* (2014) Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell Proteomics*, 13, 1690–1704.
 32. Stokes,M.P., Farnsworth,C.L., Moritz,A. *et al.* (2012) PTMScore direct: identification and quantification of peptides from critical signaling proteins by immunoaffinity enrichment coupled with LC-MS/MS. *Mol. Cell Proteomics*, 11, 187–201.

Conclusion

Cette thèse s'est attachée à proposer de nouvelles contributions dans la toxicologie, la toxicologie prédictive et dans l'évaluation des risques chimiques. Ce travail s'inscrit dans le cadre des nombreuses réglementations sur la production et la distribution des composés chimiques ayant vu le jour ces dernières années. Celles-ci incitent la communauté scientifique à développer des méthodes alternatives orientant les recherches et facilitant l'estimation des risques des molécules. Dans cette thèse, je me suis plus particulièrement attaché aux problèmes posés par les perturbateurs endocriniens (PEs).

Le défi scientifique posé par les PEs est loin d'être résolu. Néanmoins, la compréhension de leurs mécanismes d'action couplés à l'évolution des technologies en toxicologie (RNA-seq, single cell, organe-on-chip) laisse entrevoir la mise en place de nouvelles stratégies de tests (Gundert-Remy et al. 2015). Les PEs possédant de nombreuses cibles, il est nécessaire de mettre en œuvre une stratégie prenant en compte la multiplicité de ces cibles et des voies de signalisation moléculaire impliquées. Ces approches de biologie systémique tentent d'intégrer l'ensemble des informations disponibles à différents niveaux de complexité (génomique, moléculaire, cellulaire, physiologique...) afin de mettre en relation toutes ces données sous forme de réseaux et de construire plus facilement des modèles biologiques les plus proches possible des modèles *in vivo* (Gautier, Taboureau, and Audouze 2013)). Dans ce contexte, les nouvelles technologies en toxicologie constituent de puissants outils pour la caractérisation des risques des PE s'affranchissant des limites inhérentes aux tests ou offrant des approches à différents niveaux.

Les approches dites « omiques » sont des outils de choix pour comprendre les grands mécanismes d'action de la toxicité des PEs, et notamment pour aborder les études des effets à faibles doses et de mélanges (Altenburger et al. 2012). Les approches de Single Cell RNA-seq prodiguent une meilleure compréhension au niveau cellulaire des effets toxiques d'une molécule (B. Zhang et al. 2017). Elles offrent la possibilité de visualiser les signatures transcriptomiques de type cellulaire minoritaire qui normalement ne sont pas visible, car « diluées » dans les cellules du tissu. C'est notamment le cas pour le rein qui est un organe constitué de nombreux types cellulaires. L'utilisation de ces approches permet également de limiter les variations intra-individuelles inhérentes aux processus d'expérimentation. L'évolution de ces technologies ces dernières années a permis l'accès à des technologies de transcriptomiques de faible coût tel que le DGE-seq. Pour 25€, il est maintenant possible de faire séquencer un échantillon alors que le séquençage classique peut monter jusqu'à 750€ par échantillon. Cette diminution du prix impacte directement les expérimentations, offrant la possibilité de

faire séquencer plus d'échantillons, d'augmenter le nombre de répliques et ainsi augmenter la puissance statistique des analyses.

Enfin, l'utilisation de ces techniques dans le cadre de stratégies prédictives *in silico* pourrait non seulement permettre l'extrapolation espèces expérimentales/homme, mais aussi définir des profils toxicologiques, de modéliser la courbe de dose-réponse d'un composé ou encore évaluer une exposition à une substance (Suter, Babiss, and Wheeldon 2004). Les organes sur puces (ou organe-on-chip) prennent en compte les interactions complexes du vivant et ont ainsi une approche méthodologique plus pertinente (Pamies and Hartung 2017).

De nombreux programmes d'évaluation des risques présentent parmi leurs objectifs un volet basé sur le développement de telles méthodes. Outre la Commission européenne avec REACH, l' US EPA avec Tox21, d'autres sont tout aussi importants dans le domaine de la toxicologie prédictive. eTOX (Taboureau et al. 2012), SEURAT (Daston et al. 2015), PREDICT-IV (Pfaller et al. 2015) ou encore The Human Toxome Project (Bouhifd et al. 2015), pour ne citer qu'eux proposent la mise en place de protocoles dans le but d'affiner le spectre d'informations accessibles sur les composés chimiques et à terme mettre en place des outils prédictifs plus fins.

Si la plupart des outils prédictifs utilisent les QSARs ou les modèles PBPKs, de nouvelles méthodes basées sur la protéomique ou la transcriptomique voient également le jour (Crawford et al. 2017). C'est dans ce contexte que s'est inscrit mon projet de thèse et plus spécifiquement le projet ChemPSy. Visant au départ à estimer la possibilité de regrouper des pathologies entre elles en fonction de profils de transcriptomiques particuliers, l'envergure du projet a été étendue au fur et à mesure de ses avancements. Dans un premier temps, ChemPSy a permis de redémontrer à grande échelle la preuve de concept établi par Steiner et ses collaborateurs (Steiner et al. 2004) : des signatures toxicogénomiques proches sont associées à des effets délétères proches. Cependant, ChemPSy est allé au-delà en extrapolant ces résultats sur d'autres organes en utilisant une myriade de composées possédant des effets délétères hétérogènes et non centrés sur l'hépatotoxicité. Cette étape de classification a pu mettre en avant une vingtaine de molécules, dont l'effet PE est connu pour la plupart d'entre elles et dont l'effet pour cinq d'entre elles reste à démontrer lors de l'étape de validation expérimentale actuellement en cours. Enfin, si l'étape de prédiction de phénotype semble fonctionner, il sera nécessaire d'améliorer sa sensibilité par l'ajout de nouvelles données transcriptomiques, mais également par l'ajout d'autres informations structurales, pharmacocinétiques ou même métabolomiques. Dans le prolongement de ces recherches, un autre projet a vu le jour : MASSIVE ATTACK. Ce projet permettra d'agrandir de manière significative le nombre de signatures toxicogénomiques présentes dans ChemPSy. Ceci aura pour résultat

d'homogénéiser le jeu de données d'un point de vue expérimental. Tous les composés seront testés au même temps, à la même dose, avec le même nombre de répliques et ce par l'utilisation d'un protocole standardisé automatisé afin de limiter les variabilités inter-échantillons. De plus l'intégration des données de métabolomiques associées aux signatures permettra d'affiner la génération de groupe en prenant en compte ces nouvelles informations pour la classification.

Au vu de l'orientation des recherches sur les faibles doses et les mélanges, la classification sans a priori et l'association basée uniquement sur la signature transcriptomique laissent à penser que ChemPSy sera compatible avec de telles données.

Le développement de ChemPSy a mis en avant les difficultés rencontrées dans le domaine de la fouille de données et la réutilisation d'informations, plus particulièrement dans le domaine de la toxicologie. La constitution et la mise en forme des données de ChemPSy a mis plus de 6 mois. Le problème est l'hétérogénéité des données nécessitant un gros travail pour les préparer et les uniformiser (même temps, dose, nom molécule, ...). De plus les données ayant trait à la toxicogénomique sont actuellement « noyées » au sein de dépôts de données généralistes et leur récupération est loin d'être aisée. C'est pourquoi, durant la phase de constitution du jeu de données de ChemPSy, le projet TOXsIgN a été initié en parallèle. Cet espace de dépôt a pour vocation de structurer et d'ordonner les données de toxicologie issues d'analyses génomiques, moléculaires et/ou physiologiques ; ceci grâce à l'utilisation de vocabulaire contrôlé. TOXsIgN vient en complément des banques déjà mises en place, telles que diXa, CEB ou ToxDB, en donnant accès à un nouveau type d'information : les signatures toxicogénomiques. Ce dépôt compatible avec les études transgénérationnelles et les mélanges possède une architecture souple lui permettant d'intégrer de nombreux outils, mais également de s'adapter à de nombreux type de données de toxicologie. Le succès de cette ressource repose aujourd'hui principalement sur notre capacité à fournir un ensemble de données assez volumineux pour attirer la communauté scientifique afin qu'elle s'empare de ce dépôt et le fasse vivre. Il repose également sur le type d'outils mis à disposition et notamment un navigateur de génome permettra aux toxicologues de visualiser les données issues de séquençage afin de faciliter l'interprétation et la création de figures pour les publications. Un premier essai a déjà été effectué dans ce sens par la création du navigateur de génome centré autour de la reproduction : *The ReproGenomics Viewer*.

En conclusion, ChemPSy, TOXsIgN et dans une moindre mesure RGV participent à la mise en place de nouvelles méthodologies dans l'échange, le stockage, la réutilisation des données toxicologiques ainsi que dans l'évaluation des risques chimiques. Au vu du coût annuel estimé des maladies résultant de l'exposition aux PE (Trasande et al. 2015), il est facile de se rendre compte de la problématique

économique et sociale soulevée par ces molécules. Leur détection reste donc pour les autorités sanitaires une priorité favorisant les approches prédictives innovantes. Cela se comprend d'autant plus lorsque l'on constate les économies effectuées lors de l'utilisation de telles méthodes dans l'évaluation des risques sanitaires d'une molécule (Pedersen et al. 2003). Les avancées technologiques permettent, aujourd'hui, de réaliser à la fois des tests plus sensibles et spécifiques, mais également des approches intégratives plus poussées. Tous ces tests sont autant d'outils pour évaluer les PE et nécessitent d'être organisés en définissant une stratégie générale d'analyse. Dans tous les cas il est fondamental de définir comment les informations obtenues peuvent être utilisées pour identifier, évaluer et prévenir le risque chimique.

Références

- A. Bergman, J. J. Heindel, S. Jobling, K. A. Kidd, R. T. Zoeller. 2013. "State of the Science of Endocrine Disrupting Chemicals - 2012." *WHO Press*. World Health Organization. <http://www.who.int/ceh/publications/endocrine/en/>.
- Abiven, G, M.-L Raffin-Sanson, and J Bertherat. 2004. "Biochimie Des Hormones et Leurs Mécanismes D'action. Généralités et Synthèse Des Hormones Polypeptidiques." *EMC - Endocrinologie* 1 (2): 81–92. doi:10.1016/j.emcend.2004.01.003.
- Acevedo, Hernan F. 2002. "Human Chorionic Gonadotropin (hCG), the Hormone of Life and Death: A Review." *Journal of Experimental Therapeutics & Oncology* 2 (3): 133–45. <http://www.ncbi.nlm.nih.gov/pubmed/12415629>.
- Adams, N R. 1990. "Permanent Infertility in Ewes Exposed to Plant Oestrogens." *Australian Veterinary Journal* 67 (6): 197–201. <http://www.ncbi.nlm.nih.gov/pubmed/2222361>.
- AFSSA. 2009. "Évaluation Des Risques Liés À La Présence de Mycotoxines Dans Les Chaînes Alimentaires Humaine et Animale Rapport Final." <https://www.anses.fr/fr/system/files/RCCP-Ra-Mycotoxines2009.pdf>.
- Aken, Bronwen L., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Friederike Bernsdorff, Jyothish Bhai, Konstantinos Billis, et al. 2017. "Ensembl 2017." *Nucleic Acids Research* 45 (D1): D635–42. doi:10.1093/nar/gkw1104.
- Albert, O, C Desdoits-Lethimonier, L Lesné, A Legrand, F Guillé, K Bensalah, N Dejuicq-Rainsford, and B Jégou. 2013. "Paracetamol, Aspirin and Indomethacin Display Endocrine Disrupting Properties in the Adult Human Testis in Vitro." *Human Reproduction (Oxford, England)* 28 (7): 1890–98. doi:10.1093/humrep/det112.
- Albert, Océane, and Bernard Jégou. 2014. "A Critical Assessment of the Endocrine Susceptibility of the Human Testis to Phthalates from Fetal Life to Adulthood." *Human Reproduction Update* 20 (2): 231–49. doi:10.1093/humupd/dmt050.
- Altenburger, Rolf, Stefan Scholz, Mechthild Schmitt-Jansen, Wibke Busch, and Beate I. Escher. 2012. "Mixture Toxicity Revisited from a Toxicogenomic Perspective." *Environmental Science & Technology* 46 (5): 2508–22. doi:10.1021/es2038036.
- Altenhoff, A. M., N. kunca, N. Glover, C.-M. Train, A. Sueki, I. Pili ota, K. Gori, et al. 2015. "The OMA Orthology Database in 2015: Function Predictions, Better Plant Support, Synteny View and Other Improvements." *Nucleic Acids Research* 43 (D1): D240–49. doi:10.1093/nar/gku1158.
- Amberger, J. S., C. A. Bocchini, F. Schiettecatte, A. F. Scott, and A. Hamosh. 2015. "OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an Online Catalog of Human Genes and Genetic Disorders." *Nucleic Acids Research* 43 (D1): D789–98. doi:10.1093/nar/gku1205.
- Ames, B N, W E Durston, E Yamasaki, and F D Lee. 1973. "Carcinogens Are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection." *Proceedings of the National Academy of Sciences of the United States of America* 70 (8): 2281–85. <http://www.ncbi.nlm.nih.gov/pubmed/4151811>.
- Andersson, Patrik L., Jerker Fick, and Stefan Rännar. 2011. "A Multivariate Chemical Similarity Approach to Search for Drugs of Potential Environmental Concern." *Journal of Chemical Information and Modeling* 51 (8): 1788–94. doi:10.1021/ci200107b.
- Ankley, Gerald T, Kathleen M Jensen, Elizabeth A Makynen, Michael D Kahl, Joseph J

- Korte, Michael W Hornung, Tala R Henry, et al. 2003. "Effects of the Androgenic Growth Promoter 17-Beta-Trenbolone on Fecundity and Reproductive Endocrinology of the Fathead Minnow." *Environmental Toxicology and Chemistry* 22 (6): 1350–60. <http://www.ncbi.nlm.nih.gov/pubmed/12785594>.
- Ashburner, M, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, et al. 2000. "Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium." *Nature Genetics* 25 (1): 25–29. doi:10.1038/75556.
- Ashby, J, and R W Tennant. 1988. "Chemical Structure, Salmonella Mutagenicity and Extent of Carcinogenicity as Indicators of Genotoxic Carcinogenesis among 222 Chemicals Tested in Rodents by the U.S. NCI/NTP." *Mutation Research* 204 (1): 17–115. <http://www.ncbi.nlm.nih.gov/pubmed/3277047>.
- Audouze, Karine, and Olivier Taboureau. 2015. "Chemical Biology Databases: From Aggregation, Curation to Representation." *Drug Discovery Today: Technologies* 14 (July): 25–29. doi:10.1016/j.ddtec.2015.03.003.
- Axmon, Anna, Lars Rylander, Ulf Strömberg, and Lars Hagmar. 2004. "Altered Menstrual Cycles in Women with a High Dietary Intake of Persistent Organochlorine Compounds." *Chemosphere* 56 (8): 813–19. doi:10.1016/j.chemosphere.2004.03.002.
- Bajusz, Dávid, Anita Rácz, and Károly Héberger. 2015. "Why Is Tanimoto Index an Appropriate Choice for Fingerprint-Based Similarity Calculations?" *Journal of Cheminformatics* 7. Springer: 20. doi:10.1186/s13321-015-0069-3.
- Bard, Jonathan, Seung Y Rhee, and Michael Ashburner. 2005. "An Ontology for Cell Types." *Genome Biology* 6 (2): R21. doi:10.1186/gb-2005-6-2-r21.
- Barrett, Tanya, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, et al. 2013. "NCBI GEO: Archive for Functional Genomics Data Sets--Update." *Nucleic Acids Research* 41 (Database issue): D991-5. doi:10.1093/nar/gks1193.
- Baskin, L S, K Himes, and T Colborn. 2001. "Hypospadias and Endocrine Disruption: Is There a Connection?" *Environmental Health Perspectives* 109 (11). National Institute of Environmental Health Science: 1175–83. <http://www.ncbi.nlm.nih.gov/pubmed/11713004>.
- Baskin, Laurence. 2017. "What Is Hypospadias?" *Clinical Pediatrics* 56 (5): 409–18. doi:10.1177/0009922816684613.
- Benigni, Romualdo, and Cecilia Bossa. 2008. "Structure Alerts for Carcinogenicity, and the Salmonella Assay System: A Novel Insight through the Chemical Relational Databases Technology." *Mutation Research* 659 (3): 248–61. doi:10.1016/j.mrrev.2008.05.003.
- Bennett, Kristin. 2000. "Support Vector Machines: Hype or Hallelujah?" *SIGKDD Explorations*, 2–1. <https://pdfs.semanticscholar.org/04a9/11ad5010129a17340fd7b7be53ce581e75d6.pdf>.
- Béranger, Rémi, Ronan Garlantézec, Gaïd Le Maner-Idrissi, Agnès Lacroix, Florence Rouget, Jessica Trowbridge, Charline Warembourg, et al. 2017. "Prenatal Exposure to Glycol Ethers and Neurocognitive Abilities in 6-Year-Old Children: The PELAGIE Cohort Study." *Environmental Health Perspectives* 125 (4): 684–90. doi:10.1289/EHP39.
- Bergman, Jorieke E. H., Maria Loane, Martine Vrijheid, Anna Pierini, Rien J. M. Nijman, Marie-Claude Addor, Ingeborg Barisic, et al. 2015. "Epidemiology of Hypospadias in Europe: A Registry-Based Study." *World Journal of Urology* 33 (12): 2159–67. doi:10.1007/s00345-015-1507-6.

- Bertherat, J. 2004. "Biochimie Des Hormones et Leurs Mécanismes D'action. D-Récepteurs Nucléaires." *EMC - Endocrinologie* 1 (3): 133–37. doi:10.1016/j.emcend.2004.03.003.
- Bhattachar, Shobha N., Everett J. Perkins, Jeffrey S. Tan, and Lee J. Burns. 2011. "Effect of Gastric pH on the Pharmacokinetics of a Bcs Class II Compound in Dogs: Utilization of an Artificial Stomach and Duodenum Dissolution Model and Gastroplus,TM Simulations to Predict Absorption." *Journal of Pharmaceutical Sciences* 100 (11): 4756–65. doi:10.1002/jps.22669.
- Bhatti, J S, I P S Sidhu, and G K Bhatti. 2011. "Ameliorative Action of Melatonin on Oxidative Damage Induced by Atrazine Toxicity in Rat Erythrocytes." *Molecular and Cellular Biochemistry* 353 (1–2): 139–49. doi:10.1007/s11010-011-0780-y.
- Birgersdotter, Anna, Rickard Sandberg, and Ingemar Ernberg. 2005. "Gene Expression Perturbation in Vitro--a Growing Case for Three-Dimensional (3D) Culture Systems." *Seminars in Cancer Biology* 15 (5): 405–12. doi:10.1016/j.semcancer.2005.06.009.
- Blankenberg, Daniel, Gregory Von Kuster, Nathaniel Coraor, Guruprasad Ananda, Ross Lazarus, Mary Mangan, Anton Nekrutenko, and James Taylor. 2010. "Galaxy: A Web-Based Genome Analysis Tool for Experimentalists." *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* Chapter 19 (January): Unit 19.10.1-21. doi:10.1002/0471142727.mb1910s89.
- Blount, B C, M J Silva, S P Caudill, L L Needham, J L Pirkle, E J Sampson, G W Lucier, R J Jackson, and J W Brock. 2000. "Levels of Seven Urinary Phthalate Metabolites in a Human Reference Population." *Environmental Health Perspectives* 108 (10): 979–82. <http://www.ncbi.nlm.nih.gov/pubmed/11049818>.
- Boué, Stephen M., Thomas E. Wiese, Suzanne Nehls, Matthew E. Burow, Steven Elliott, Carol H. Carter-Wientjes, Betty Y. Shih, John A. McLachlan, and Thomas E. Cleveland. 2003. "Evaluation of the Estrogenic Effects of Legume Extracts Containing Phytoestrogens." *Journal of Agricultural and Food Chemistry* 51 (8): 2193–99. doi:10.1021/jf021114s.
- Bouhifd, Mounir, Melvin E Andersen, Christina Baghdikian, Kim Boekelheide, Kevin M Crofton, Albert J Fornace, Andre Kleensang, et al. 2015. "The Human Toxome Project." *ALTEX* 32 (2): 112–24. doi:http://dx.doi.org/10.14573/altex.1502091.
- Boyle, C. A., S. Boulet, L. A. Schieve, R. A. Cohen, S. J. Blumberg, M. Yeargin-Allsopp, S. Visser, and M. D. Kogan. 2011. "Trends in the Prevalence of Developmental Disabilities in US Children, 1997-2008." *PEDIATRICS* 127 (6): 1034–42. doi:10.1542/peds.2010-2989.
- Bras, Yvan Le, Aurélien Roult, Cyril Monjeaud, Mathieu Bahin, Olivier Quenez, Claudia Heriveau, Anthony Bretaudeau, Olivier Sallou, and Olivier Collin. 2013. "Towards a Life Sciences Virtual Research Environment An E-Science Initiative in Western France Vers Un Environnement Virtuel de Recherche En Sciences de La Vie." <https://www.e-biogenouest.org/resources/128>. https://www.e-biogenouest.org/resources/129/download/jobim_YLeBras_2013.pdf.
- Breslin, Susan, and Lorraine O'Driscoll. 2013. "Three-Dimensional Cell Culture: The Missing Link in Drug Discovery." *Drug Discovery Today* 18 (5–6): 240–49. doi:10.1016/j.drudis.2012.10.003.
- Brinkman, Ryan R, Mélanie Courtot, Dirk Derom, Jennifer M Fostel, Yongqun He, Phillip Lord, James Malone, et al. 2010. "Modeling Biomedical Experimental Processes with OBI." *Journal of Biomedical Semantics* 1 Suppl 1 (January): S7. doi:10.1186/2041-1480-

1-S1-S7.

- Britto, Ramona, Olivier Sallou, Olivier Collin, Grégoire Michaux, Michael Primig, and Frédéric Chalmel. 2012. "GPSy: A Cross-Species Gene Prioritization System for Conserved Biological Processes--Application in Male Gamete Development." *Nucleic Acids Research* 40 (Web Server issue): W458-65. doi:10.1093/nar/gks380.
- Brown, Ronald P., Michael D. Delp, Stan L. Lindstedt, Lorenz R. Rhomberg, and Robert P. Beliles. 1997. "Physiological Parameter Values for Physiologically Based Pharmacokinetic Models." *Toxicology and Industrial Health* 13 (4): 407-84. doi:10.1177/074823379701300401.
- Bulera, S J, S M Eddy, E Ferguson, T A Jatkoa, J F Reindel, M R Bleavins, and F A De La Iglesia. 2001. "RNA Expression in the Early Characterization of Hepatotoxins in Wistar Rats by High-Density DNA Microarrays." *Hepatology (Baltimore, Md.)* 33 (5): 1239-58. doi:10.1053/jhep.2001.23560.
- Buluş, Ayşe Derya, Ali Aşci, Pinar Erkekoglu, Aylin Balci, Nesibe Andiran, and Belma Koçer-Gümüşel. 2016. "The Evaluation of Possible Role of Endocrine Disruptors in Central and Peripheral Precocious Puberty." *Toxicology Mechanisms and Methods* 26 (7): 493-500. doi:10.3109/15376516.2016.1158894.
- Calafat, Antonia M, Xiaoyun Ye, Lee-Yang Wong, John A Reidy, and Larry L Needham. 2008. "Exposure of the U.S. Population to Bisphenol A and 4-Tertiary-Octylphenol: 2003-2004." *Environmental Health Perspectives* 116 (1): 39-44. doi:10.1289/ehp.10753.
- Cannon, S B, J M Veazey, R S Jackson, V W Burse, C Hayes, W E Straub, P J Landrigan, and J A Liddle. 1978. "Epidemic Kepone Poisoning in Chemical Workers." *American Journal of Epidemiology* 107 (6): 529-37. <http://www.ncbi.nlm.nih.gov/pubmed/78669>.
- Cariello, Neal F, John D Wilson, Ben H Britt, David J Wedd, Brian Burlinson, and Vijay Gombar. 2002. "Comparison of the Computer Programs DEREK and TOPKAT to Predict Bacterial Mutagenicity. Deductive Estimate of Risk from Existing Knowledge. Toxicity Prediction by Komputer Assisted Technology." *Mutagenesis* 17 (4): 321-29. <http://www.ncbi.nlm.nih.gov/pubmed/12110629>.
- Carlsen, E, A Giwercman, N Keiding, and N E Skakkebaek. 1992. "Evidence for Decreasing Quality of Semen during Past 50 Years." *BMJ (Clinical Research Ed.)* 305 (6854): 609-13. <http://www.ncbi.nlm.nih.gov/pubmed/1393072>.
- Carlsen, Lars, Bulat N Kenessov, and Svetlana Ye Batyrbekova. 2008. "A QSAR/QSTR Study on the Environmental Health Impact by the Rocket Fuel 1,1-Dimethyl Hydrazine and Its Transformation Products." *Environmental Health Insights* 1 (July): 11-20. <http://www.ncbi.nlm.nih.gov/pubmed/21572843>.
- Carson, Rachel. 1962. *Silent Spring*. Houghton Mifflin.
- Case, David A., Thomas E. Cheatham, Tom Darden, Holger Gohlke, Ray Luo, Kenneth M. Merz, Alexey Onufriev, Carlos Simmerling, Bing Wang, and Robert J. Woods. 2005. "The Amber Biomolecular Simulation Programs." *Journal of Computational Chemistry* 26 (16): 1668-88. doi:10.1002/jcc.20290.
- Cassone, V M, W S Warren, D S Brooks, and J Lu. 1993. "Melatonin, the Pineal Gland, and Circadian Rhythms." *Journal of Biological Rhythms* 8 Suppl: S73-81. <http://www.ncbi.nlm.nih.gov/pubmed/8274765>.
- Catena, Cristiana, GianLuca Colussi, Francesca Nait, Flavia Martinis, Francesca Pezzutto, and Leonardo A Sechi. 2014. "Aldosterone and the Heart: Still an Unresolved Issue?" *Frontiers in Endocrinology* 5 (October): 168. doi:10.3389/fendo.2014.00168.

- Chao, H-R, S-L Wang, L-Y Lin, W-J Lee, and O Pöpke. 2007. "Placental Transfer of Polychlorinated Dibenzo-P-Dioxins, Dibenzofurans, and Biphenyls in Taiwanese Mothers in Relation to Menstrual Cycle Characteristics." *Food and Chemical Toxicology : An International Journal Published for the British Industrial Biological Research Association* 45 (2): 259–65. doi:10.1016/j.fct.2006.07.032.
- Cheng, L, and L Li. 2016. "Systematic Quality Control Analysis of LINCS Data." *CPT: Pharmacometrics & Systems Pharmacology* 5 (11): 588–98. doi:10.1002/psp4.12107.
- Chevrier, Cécile, Gwendolina Limon, Christine Monfort, Florence Rouget, Ronan Garlantézec, Claire Petit, Gaël Durand, and Sylvaine Cordier. 2011. "Urinary Biomarkers of Prenatal Atrazine Exposure and Adverse Birth Outcomes in the PELAGIE Birth Cohort." *Environmental Health Perspectives* 119 (7): 1034–41. doi:10.1289/ehp.1002775.
- Chia, Victoria M, Sabah M Quraishi, Susan S Devesa, Mark P Purdue, Michael B Cook, and Katherine A McGlynn. 2010. "International Trends in the Incidence of Testicular Cancer, 1973-2002." *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology* 19 (5): 1151–59. doi:10.1158/1055-9965.EPI-10-0031.
- Cho, Soohee, and Jeong-Yeol Yoon. 2017. "Organ-on-a-Chip for Assessing Environmental Toxicants." *Current Opinion in Biotechnology* 45 (June): 34–42. doi:10.1016/j.copbio.2016.11.019.
- Choo, Siew Woh, Mia Yang Ang, Hanieh Fouladi, Shi Yang Tan, Cheuk Chuen Siow, Naresh V R Mutha, Hamed Heydari, et al. 2014. "HelicoBase: A Helicobacter Genomic Resource and Analysis Platform." *BMC Genomics* 15 (1): 600. doi:10.1186/1471-2164-15-600.
- Choo, Siew Woh, Hamed Heydari, Tze King Tan, Cheuk Chuen Siow, Ching Yew Beh, Wei Yee Wee, Naresh V R Mutha, Guat Jah Wong, Mia Yang Ang, and Amir Hessam Yazdi. 2014. "VibrioBase: A Model for Next-Generation Genome and Annotation Database Development." *TheScientificWorldJournal* 2014 (January): 569324. doi:10.1155/2014/569324.
- Chris Fraley, Adrian E. Raftery, T. Brendan Murphy, and Luca Scrucca. 2012. "Mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation." <http://my.ilstu.edu/~mxu2/mat456/mcluster.pdf>.
- Christen, Markus, Philippe H. Hünenberger, Dirk Bakowies, Riccardo Baron, Roland Bürgi, Daan P. Geerke, Tim N. Heinz, et al. 2005. "The GROMOS Software for Biomolecular Simulation: GROMOS05." *Journal of Computational Chemistry* 26 (16): 1719–51. doi:10.1002/jcc.20303.
- Clermont, Yves. 1963. "The Cycle of the Seminiferous Epithelium in Man." *American Journal of Anatomy* 112 (1): 35–51. doi:10.1002/aja.1001120103.
- Colborn, Theo, Dianne Dumanoski, and John Peterson Myers. 1996. *Our Stolen Future : Are We Threatening Our Fertility, Intelligence, and Survival? : A Scientific Detective Story*. <http://www.ourstolenfuture.org/aboutosf.htm#>.
- Collins, Francis S, and Lawrence A Tabak. 2014. "Policy: NIH Plans to Enhance Reproducibility." *Nature* 505 (7485): 612–13. <http://www.ncbi.nlm.nih.gov/pubmed/24482835>.
- Commission, European. 1996. "EUROPEAN WORKSHOP ON THE IMPACT OF ENDOCRINE DISRUPTERS ON HUMAN HEALTH AND WILDLIFE."

- http://www.iehconsulting.co.uk/IEH_Consulting/IEHCPubs/EndocrineDisrupters/WEYB RIDGE.pdf.
- Conley, A, and M Hinshelwood. 2001. "Mammalian Aromatases." *Reproduction (Cambridge, England)* 121 (5): 685–95. <http://www.ncbi.nlm.nih.gov/pubmed/11427156>.
- Contrera, Joseph F., Naomi L. Kruhlak, Edwin J. Matthews, and R. Daniel Benz. 2007. "Comparison of MC4PC and MDL-QSAR Rodent Carcinogenicity Predictions and the Enhancement of Predictive Performance by Combining QSAR Models." *Regulatory Toxicology and Pharmacology* 49 (3): 172–82. doi:10.1016/j.yrtph.2007.07.001.
- Cooper, Amber R., Valerie L. Baker, Evelina W. Sterling, Mary E. Ryan, Teresa K. Woodruff, and Lawrence M. Nelson. 2011. "The Time Is Now for a New Approach to Primary Ovarian Insufficiency." *Fertility and Sterility* 95 (6): 1890–97. doi:10.1016/j.fertnstert.2010.01.016.
- Counis, Raymond, Jean-Noël Laverrière, Ghislaine Garrel, Christian Bleux, Joëlle Cohen-Tannoudji, Yannick Lerrant, Marie-Laure Kottler, and Solange Magre. 2005. "Gonadotropin-Releasing Hormone and the Control of Gonadotrope Function." *Reproduction Nutrition Development* 45 (3): 243–54. doi:10.1051/rnd:2005017.
- Crawford, Sarah E, Thomas Hartung, Henner Hollert, Björn Mathes, Bennard van Ravenzwaay, Thomas Steger-Hartmann, Christoph Studer, and Harald F Krug. 2017. "Green Toxicology: A Strategy for Sustainable Chemical and Material Development." *Environmental Sciences Europe* 29 (1). Springer: 16. doi:10.1186/s12302-017-0115-z.
- Dai, Manhong, Pinglang Wang, Andrew D. Boyd, Georgi Kostov, Brian Athey, Edward G. Jones, William E. Bunney, et al. 2005. "Evolving Gene/transcript Definitions Significantly Alter the Interpretation of GeneChip Data." *Nucleic Acids Research* 33 (20). Oxford University Press: e175. doi:10.1093/nar/gni179.
- Darde, Thomas a., Olivier Sallou, Emmanuelle Becker, Bertrand Evrard, Cyril Monjeaud, Yvan Le Bras, Bernard Jégou, et al. 2015. "The ReproGenomics Viewer: An Integrative Cross-Species Toolbox for the Reproductive Science Community." *Nucleic Acids Research* 43 (W1): 1–8. doi:10.1093/nar/gkv345.
- Daston, George, Derek J. Knight, Michael Schwarz, Tilman Gocht, Russell S. Thomas, Catherine Mahony, and Maurice Whelan. 2015. "SEURAT: Safety Evaluation Ultimately Replacing Animal Testing—Recommendations for Future Research in the Field of Predictive Toxicology." *Archives of Toxicology* 89 (1): 15–23. doi:10.1007/s00204-014-1421-5.
- Davies, Mark, Michał Nowotka, George Papadatos, Nathan Dedman, Anna Gaulton, Francis Atkinson, Louisa Bellis, and John P Overington. 2015. "ChEMBL Web Services: Streamlining Access to Drug Discovery Data and Utilities." *Nucleic Acids Research* 43 (W1): W612-20. doi:10.1093/nar/gkv352.
- Davis, Allan Peter, Cynthia J Grondin, Kelley Lennon-Hopkins, Cynthia Saraceni-Richards, Daniela Sciaky, Benjamin L King, Thomas C Wieggers, and Carolyn J Mattingly. 2015. "The Comparative Toxicogenomics Database's 10th Year Anniversary: Update 2015." *Nucleic Acids Research* 43 (Database issue): D914-20. doi:10.1093/nar/gku935.
- Davis, J T, and D E Ong. 1995. "Retinol Processing by the Peritubular Cell from Rat Testis." *Biology of Reproduction* 52 (2): 356–64. <http://www.ncbi.nlm.nih.gov/pubmed/7711204>.
- Deeb, Omar, and Mohammad Goodarzi. 2012. "In Silico Quantitative Structure Toxicity Relationship of Chemical Compounds: Some Case Studies." *Current Drug Safety* 7 (4): 289–97. <http://www.ncbi.nlm.nih.gov/pubmed/23062241>.

- Degtyarenko, Kirill, Janna Hastings, Paula de Matos, and Marcus Ennis. 2009. "ChEBI: An Open Bioinformatics and Cheminformatics Resource." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* Chapter 14 (June): Unit 14.9. doi:10.1002/0471250953.bi1409s26.
- Desdoits-Lethimonier, C., L. Lesné, P. Gaudriault, D. Zalko, J.P. Antignac, Y. Deceuninck, C. Platel, N. Dejucq-Rainsford, S. Mazaud-Guittot, and B. J?gou. 2017. "Parallel Assessment of the Effects of Bisphenol A and Several of Its Analogs on the Adult Human Testis." *Human Reproduction* 32 (7): 1465–73. doi:10.1093/humrep/dex093.
- Desvergne, B?atrice, and Walter Wahli. 1999. "Peroxisome Proliferator-Activated Receptors: Nuclear Control of Metabolism." *Endocrine Reviews* 20 (5): 649–88. doi:10.1210/edrv.20.5.0380.
- Dewailly, E, P Ayotte, S Bruneau, C Laliberté, D C Muir, and R J Norstrom. 1993. "Inuit Exposure to Organochlorines through the Aquatic Food Chain in Arctic Québec." *Environmental Health Perspectives* 101 (7): 618–20. <http://www.ncbi.nlm.nih.gov/pubmed/8143594>.
- DiMasi, Joseph A., and Henry G. Grabowski. 2007. "Economics of New Oncology Drug Development." *Journal of Clinical Oncology* 25 (2): 209–16. doi:10.1200/JCO.2006.09.0803.
- Dimitrov, S., and O. Mekenyan. 2010. "An Introduction to Read-Across for the Prediction of the Effects of Chemicals." In , 372–84. doi:10.1039/9781849732093-00372.
- Dingemans, Milou M L, Martin van den Berg, and Remco H S Westerink. 2011. "Neurotoxicity of Brominated Flame Retardants: (In)direct Effects of Parent and Hydroxylated Polybrominated Diphenyl Ethers on the (Developing) Nervous System." *Environmental Health Perspectives* 119 (7): 900–907. doi:10.1289/ehp.1003035.
- Donlin, Maureen J. 2009. "Using the Generic Genome Browser (GBrowse)." *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* Chapter 9 (December): Unit 9.9. doi:10.1002/0471250953.bi0909s28.
- Dorfman, Ralph I., T. F. Gallagher, and F. C. Koch. 1935. "The Nature of the Estrogenic Substance in Human Male Urine and Bull Testis." *Endocrinology* 19 (1). Oxford University Press: 33–41. doi:10.1210/endo-19-1-33.
- Dragin, Nadine, Jacky Bismuth, Géraldine Cizeron-Clairac, Maria Grazia Biferi, Claire Berthault, Alain Serraf, Rémi Nottin, et al. 2016. "Estrogen-Mediated Downregulation of AIRE Influences Sexual Dimorphism in Autoimmune Diseases." *The Journal of Clinical Investigation* 126 (4): 1525–37. doi:10.1172/JCI81894.
- Driesche, Sander van den, Joni Macdonald, Richard A. Anderson, Zoe C. Johnston, Tarini Chetty, Lee B. Smith, Chris McKinnell, et al. 2015. "Prolonged Exposure to Acetaminophen Reduces Testosterone Production by the Human Fetal Testis in a Xenograft Model." *Science Translational Medicine* 7 (288). <http://stm.sciencemag.org.gate2.inist.fr/content/7/288/288ra80.full>.
- Drucker, D, M C Eggo, I E Salit, and G N Burrow. 1984. "Ethionamide-Induced Goitrous Hypothyroidism." *Annals of Internal Medicine* 100 (6): 837–39. <http://www.ncbi.nlm.nih.gov/pubmed/6721300>.
- Dupuis, Gilles, and Claude Benezra. 1983. "Allergic Contact Dermatitis to Simple Chemicals: A Molecular Approach." *Journal of the American Academy of Dermatology* 8 (4). Elsevier: 575. doi:10.1016/S0190-9622(83)80070-X.
- Edgar, Ron, Michael Domrachev, and Alex E Lash. 2002. "Gene Expression Omnibus: NCBI

- Gene Expression and Hybridization Array Data Repository.” *Nucleic Acids Research* 30 (1): 207–10.
<http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=99122&tool=pmcentrez&rendertype=abstract>.
- Eladak, Soria, Tiphany Grisin, Delphine Moison, Marie-Justine Guerquin, Thierry N’Tumba-Byn, Stéphanie Pozzi-Gaudin, Alexandra Benachi, Gabriel Livera, Virginie Rouiller-Fabre, and René Habert. 2015. “A New Chapter in the Bisphenol A Story: Bisphenol S and Bisphenol F Are Not Safe Alternatives to This Compound.” *Fertility and Sterility* 103 (1): 11–21. doi:10.1016/j.fertnstert.2014.11.005.
- Eltyeb, Safaa, and Naomie Salim. 2014. “Chemical Named Entities Recognition: A Review on Approaches and Applications.” *Journal of Cheminformatics* 6 (1): 17. doi:10.1186/1758-2946-6-17.
- Enoch, S J, K. R. Przybylak, and M. T. D. Cronin. 2013. “Category Formation Case Studies.” In , 127–55. doi:10.1039/9781849734400-00127.
- Escande, Aurélie, Arnaud Pillon, Nadège Servant, Jean-Pierre Cravedi, Fernando Larrea, Peter Muhn, Jean-Claude Nicolas, Vincent Cavaillès, and Patrick Balaguer. 2006. “Evaluation of Ligand Selectivity Using Reporter Cell Lines Stably Expressing Estrogen Receptor Alpha or Beta.” *Biochemical Pharmacology* 71 (10): 1459–69. doi:10.1016/j.bcp.2006.02.002.
- Eskenazi, Brenda, Jonathan Chevrier, Lisa Goldman Rosas, Henry A Anderson, Maria S Bornman, Henk Bouwman, Aimin Chen, et al. 2009. “The Pine River Statement: Human Health Consequences of DDT Use.” *Environmental Health Perspectives* 117 (9): 1359–67. doi:10.1289/ehp.11748.
- Faroon, O, S Kueberuwa, L Smith, and C DeRosa. 1995. “ATSDR Evaluation of Health Effects of Chemicals. II. Mirex and Chlordecone: Health Effects, Toxicokinetics, Human Exposure, and Environmental Fate.” *Toxicology and Industrial Health* 11 (6): 1–203. doi:10.1177/074823379501100601.
- Farr, S L, G S Cooper, J Cai, D A Savitz, and D P Sandler. 2004. “Pesticide Use and Menstrual Cycle Characteristics among Premenopausal Women in the Agricultural Health Study.” *American Journal of Epidemiology* 160 (12): 1194–1204. doi:10.1093/aje/.
- Ferlin, Alberto, Anastasia Pepe, Lisa Ganesello, Andrea Garolla, Shu Feng, Sandro Giannini, Manuela Zaccolo, et al. 2008. “Mutations in the Insulin-Like Factor 3 Receptor Are Associated With Osteoporosis.” *Journal of Bone and Mineral Research* 23 (5): 683–93. doi:10.1359/jbmr.080204.
- Fisch, Harry, Grace Hyun, and Terry W. Hensle. 2010. “Rising Hypospadias Rates: Disproving a Myth.” *Journal of Pediatric Urology* 6 (1): 37–39. doi:10.1016/j.jpurol.2009.05.005.
- Franchimont, P, S Chari, and A Demoulin. 1975. “Hypothalamus-Pituitary-Testis Interaction.” *Journal of Reproduction and Fertility* 44 (2): 335–50.
<http://www.ncbi.nlm.nih.gov/pubmed/1099197>.
- Freese, Nowlan H., David C. Norris, and Ann E. Loraine. 2016. “Integrated Genome Browser: Visual Analytics Platform for Genomics.” *Bioinformatics* 32 (14). Oxford University Press: 2089–95. doi:10.1093/bioinformatics/btw069.
- Freire, Carmen, and Sergio Koifman. 2013. “Pesticides, Depression and Suicide: A Systematic Review of the Epidemiological Evidence.” *International Journal of Hygiene and*

- Environmental Health* 216 (4): 445–60. doi:10.1016/j.ijheh.2012.12.003.
- Gammon, Derek W, Charles N Aldous, Wesley C Carr, James R Sanborn, and Keith F Pfeifer. 2005. “A Risk Assessment of Atrazine Use in California: Human Health and Ecological Aspects.” *Pest Management Science* 61 (4): 331–55. doi:10.1002/ps.1000.
- Ganter, Brigitte, Ronald D Snyder, Donald N Halbert, and May D Lee. 2006. “Toxicogenomics in Drug Discovery and Development: Mechanistic Analysis of Compound/class-Dependent Effects Using the DrugMatrix Database.” *Pharmacogenomics* 7 (7): 1025–44. doi:10.2217/14622416.7.7.1025.
- Gautier, Laurent, Olivier Taboureau, and Karine Audouze. 2013. “The Effect of Network Biology on Drug Toxicology.” *Expert Opinion on Drug Metabolism & Toxicology* 9 (11): 1409–18. doi:10.1517/17425255.2013.820704.
- Gazouli, Maria, Zhi-Xing Yao, Nouredine Boujrad, J. Christopher Corton, Martine Culty, and Vassilios Papadopoulos. 2002. “Effect of Peroxisome Proliferators on Leydig Cell Peripheral-Type Benzodiazepine Receptor Gene Expression, Hormone-Stimulated Cholesterol Transport, and Steroidogenesis: Role of the Peroxisome Proliferator-Activator Receptor Alpha.” *Endocrinology* 143 (7): 2571–83. doi:10.1210/endo.143.7.8895.
- Gely-Pernot, Aurore, Chunxiang Hao, Emmanuelle Becker, Igor Stuparevic, Christine Kervarrec, Frédéric Chalmel, Michael Primig, Bernard Jégou, and Fatima Smagulova. 2015. “The Epigenetic Processes of Meiosis in Male Mice Are Broadly Affected by the Widely Used Herbicide Atrazine.” *BMC Genomics* 16 (1): 885. doi:10.1186/s12864-015-2095-y.
- Gharagozlou, Faramarz, Reza Youssefi, Mehdi Voijani, Vahid Akbarinejad, and Ghazaleh Rafiee. 2016. “Androgen Receptor Blockade Using Flutamide Skewed Sex Ratio of Litters in Mice.” *Veterinary Research Forum : An International Quarterly Journal* 7 (2): 169–72. <http://www.ncbi.nlm.nih.gov/pubmed/27482363>.
- Ghosh, Somiranjana, Partha S. Mitra, Christopher A. Loffredo, Tomas Trnovec, Lubica Murinova, Eva Sovcikova, Svetlana Ghimbovski, Shizhu Zang, Eric P. Hoffman, and Sisir K. Dutta. 2015. “Transcriptional Profiling and Biological Pathway Analysis of Human Equivalence PCB Exposure in Vitro: Indicator of Disease and Disorder Development in Humans.” *Environmental Research* 138 (April): 202–16. doi:10.1016/j.envres.2014.12.031.
- Giardine, Belinda, Cathy Riemer, Ross C Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. “Galaxy: A Platform for Interactive Large-Scale Genome Analysis.” *Genome Research* 15 (10): 1451–55. doi:10.1101/gr.4086505.
- Gilbert, Mary E, Joanne Rovet, Zupei Chen, and Noriyuki Koibuchi. 2012. “Developmental Thyroid Hormone Disruption: Prevalence, Environmental Contaminants and Neurodevelopmental Consequences.” *Neurotoxicology* 33 (4): 842–52. doi:10.1016/j.neuro.2011.11.005.
- Gini, Giuseppina. 2016. “QSAR Methods.” In *Methods in Molecular Biology (Clifton, N.J.)*, 1425:1–20. doi:10.1007/978-1-4939-3609-0_1.
- Goecks, Jeremy, Nate Coraor, Anton Nekrutenko, James Taylor, The Galaxy Team, Anton Nekrutenko, and James Taylor. 2012. “NGS Analyses by Visualization with Trackster.” *Nature Biotechnology* 30 (11): 1036–39. doi:10.1038/nbt.2404.
- Gray, L E, J Ostby, J Furr, M Price, D N Veeramachaneni, and L Parks. 2000. “Perinatal Exposure to the Phthalates DEHP, BBP, and DINP, but Not DEP, DMP, or DOTP, Alters

- Sexual Differentiation of the Male Rat.” *Toxicological Sciences : An Official Journal of the Society of Toxicology* 58 (2): 350–65.
<http://www.ncbi.nlm.nih.gov/pubmed/11099647>.
- Gray, L, W R Kelce, E Monosson, J S Ostby, and L S Birnbaum. 1995. “Exposure to TCDD During Development Permanently Alters Reproductive Function in Male Long-Evans Rats and Hamsters: Reduced Ejaculated and Epididymal Sperm Numbers and Sex Accessory Gland Weights in Offspring with Normal Androgenic Status.” *Toxicology and Applied Pharmacology* 131 (1): 108–18. doi:10.1006/taap.1995.1052.
- Gremse, Marion, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. 2011. “The BRENDA Tissue Ontology (BTO): The First All-Integrating Ontology of All Organisms for Enzyme Sources.” *Nucleic Acids Research* 39 (Database issue): D507–13. doi:10.1093/nar/gkq968.
- Grube, Arthur, David Donaldson, Timothy Kiely, and La Wu. 2006. “Pesticide Industry Sales and Usage Report: 2006 and 2007 Market Estimates.”
https://swap.stanford.edu/20140417081610/http://www.epa.gov/opp00001/pestsales/07pestsales/market_estimates2007.pdf.
- Grün, Felix, and Bruce Blumberg. 2007. “Perturbed Nuclear Receptor Signaling by Environmental Obesogens as Emerging Factors in the Obesity Crisis.” *Reviews in Endocrine & Metabolic Disorders* 8 (2): 161–71. doi:10.1007/s11154-007-9049-x.
- . 2009. “Minireview: The Case for Obesogens.” *Molecular Endocrinology (Baltimore, Md.)* 23 (8): 1127–34. doi:10.1210/me.2008-0485.
- Grunewald, Karsten, Wido Schmidt, Christiana Unger, and Gudrun Hanschmann. 2001. “Behavior of Glyphosate and Aminomethylphosphonic Acid (AMPA) in Soils and Water of Reservoir Radeburg II Catchment (Saxony/Germany).” *Journal of Plant Nutrition and Soil Science* 164 (1). WILEY-VCH Verlag GmbH: 65–70. doi:10.1002/1522-2624(200102)164:1<65::AID-JPLN65>3.0.CO;2-G.
- Guillette, L J, T S Gross, G R Masson, J M Matter, H F Percival, and A R Woodward. 1994. “Developmental Abnormalities of the Gonad and Abnormal Sex Hormone Concentrations in Juvenile Alligators from Contaminated and Control Lakes in Florida.” *Environmental Health Perspectives* 102 (8): 680–88.
<http://www.ncbi.nlm.nih.gov/pubmed/7895709>.
- Gundert-Remy, U, H Barth, A Bürkle, G H Degen, and R Landsiedel. 2015. “Toxicology: A Discipline in Need of Academic Anchoring--the Point of View of the German Society of Toxicology.” *Archives of Toxicology* 89 (10). Springer: 1881–93. doi:10.1007/s00204-015-1577-7.
- Gurney, Jason K., Katherine A. McGlynn, James Stanley, Tony Merriman, Virginia Signal, Caroline Shaw, Richard Edwards, Lorenzo Richiardi, John Hutson, and Diana Sarfati. 2017. “Risk Factors for Cryptorchidism.” *Nature Reviews Urology*, June.
doi:10.1038/nrurol.2017.90.
- Hackenberg, Michael, Guillermo Barturen, and José L Oliver. 2011. “NGSmethDB: A Database for next-Generation Sequencing Single-Cytosine-Resolution DNA Methylation Data.” *Nucleic Acids Research* 39 (Database issue): D75–9. doi:10.1093/nar/gkq942.
- Haddad, S, M Pelekis, and K Krishnan. 1996. “A Methodology for Solving Physiologically Based Pharmacokinetic Models without the Use of Simulation Softwares.” *Toxicology Letters* 85 (2): 113–26. <http://www.ncbi.nlm.nih.gov/pubmed/8650694>.
- Hamadeh, Hisham K, Pierre R Bushel, Supriya Jayadev, Olimpia DiSorbo, Lee Bennett,

- Leping Li, Raymond Tennant, et al. 2002. "Prediction of Compound Signature Using High Density Gene Expression Profiling." *Toxicological Sciences : An Official Journal of the Society of Toxicology* 67 (2): 232–40.
<http://www.ncbi.nlm.nih.gov/pubmed/12011482>.
- Hamadeh, Hisham K, Pierre R Bushel, Supriya Jayadev, Karla Martin, Olimpia DiSorbo, Stella Sieber, Lee Bennett, et al. 2002. "Gene Expression Analysis Reveals Chemical-Specific Profiles." *Toxicological Sciences : An Official Journal of the Society of Toxicology* 67 (2): 219–31. <http://www.ncbi.nlm.nih.gov/pubmed/12011481>.
- Hara, Shuichiro, Toshifumi Takahashi, Mitsuyoshi Amita, Hideki Igarashi, Seiji Tsutsumi, and Hirohisa Kurachi. 2011. "Bezafibrate Restores the Inhibition of FSH-Induced Follicular Development and Steroidogenesis by Tumor Necrosis Factor-Alpha Through Peroxisome Proliferator-Activated Receptor-Gamma Pathway in an In Vitro Mouse Preantral Follicle Culture" 85 (5): 895–906. doi:10.1095/biolreprod.111.090738.
- Harrow, Jennifer L., Charles A. Steward, Adam Frankish, James G. Gilbert, Jose M. Gonzalez, Jane E. Loveland, Jonathan Mudge, et al. 2014. "The Vertebrate Genome Annotation Browser 10 Years on." *Nucleic Acids Research* 42 (D1): D771–79. doi:10.1093/nar/gkt1241.
- Hartung, Thomas, and Sebastian Hoffmann. 2009. "Food for Thought ... on in Silico Methods in Toxicology." *ALTEX* 26 (3): 155–66. <http://www.ncbi.nlm.nih.gov/pubmed/19907903>.
- Harvey, Philip W, and David J Everett. 2006. "Regulation of Endocrine-Disrupting Chemicals: Critical Overview and Deficiencies in Toxicology and Risk Assessment for Human Health." *Best Practice & Research. Clinical Endocrinology & Metabolism* 20 (1): 145–65. doi:10.1016/j.beem.2005.09.008.
- Hastings, Janna, Paula de Matos, Adriano Dekker, Marcus Ennis, Bhavana Harsha, Namrata Kale, Venkatesh Muthukrishnan, et al. 2013. "The ChEBI Reference Database and Ontology for Biologically Relevant Chemistry: Enhancements for 2013." *Nucleic Acids Research* 41 (Database issue): D456–63. doi:10.1093/nar/gks1146.
- Hauser, Russ, Larisa Altshul, Zuying Chen, Louise Ryan, James Overstreet, Isaac Schiff, and David C Christiani. 2002. "Environmental Organochlorines and Semen Quality: Results of a Pilot Study." *Environmental Health Perspectives* 110 (3): 229–33.
<http://www.ncbi.nlm.nih.gov/pubmed/11882472>.
- Hauser, Russ, John D. Meeker, Susan Duty, Manori J. Silva, and Antonia M. Calafat. 2006. "Altered Semen Quality in Relation to Urinary Concentrations of Phthalate Monoester and Oxidative Metabolites." *Epidemiology* 17 (6): 682–91. doi:10.1097/01.ede.0000235996.89953.d7.
- Hayes, Tyrone B, Atif Collins, Melissa Lee, Magdalena Mendoza, Nigel Noriega, A Ali Stuart, and Aaron Vonk. 2002. "Hermaphroditic, Demasculinized Frogs after Exposure to the Herbicide Atrazine at Low Ecologically Relevant Doses." *Proceedings of the National Academy of Sciences of the United States of America* 99 (8): 5476–80. doi:10.1073/pnas.082121499.
- Hayes, Tyrone B, Vicky Khoury, Anne Narayan, Mariam Nazir, Andrew Park, Travis Brown, Lillian Adame, et al. 2010. "Atrazine Induces Complete Feminization and Chemical Castration in Male African Clawed Frogs (*Xenopus Laevis*)." *Proceedings of the National Academy of Sciences of the United States of America* 107 (10): 4612–17. doi:10.1073/pnas.0909519107.
- Hayes, Tyrone, Kelly Haston, Mable Tsui, Anhthu Hoang, Cathryn Haeffele, and Aaron

- Vonk. 2003. "Atrazine-Induced Hermaphroditism at 0.1 Ppb in American Leopard Frogs (*Rana pipiens*): Laboratory and Field Evidence." *Environmental Health Perspectives* 111 (4): 568–75. <http://www.ncbi.nlm.nih.gov/pubmed/12676617>.
- Hecker, Markus, John L Newsted, Margaret B Murphy, Eric B Higley, Paul D Jones, Rudolf Wu, and John P Giesy. 2006. "Human Adrenocarcinoma (H295R) Cells for Rapid in Vitro Determination of Effects on Steroidogenesis: Hormone Production." *Toxicology and Applied Pharmacology* 217 (1): 114–24. doi:10.1016/j.taap.2006.07.007.
- Hendrickx, Diana M., Hugo J.W.L. W. L. Aerts, Florian Caiment, Dominic Clark, Timothy M.D. D. Ebbels, Chris T. Evelo, Hans Gmuender, et al. 2015. "diXa: A Data Infrastructure for Chemical Safety Assessment." *Bioinformatics* 31 (9): 1505–7. doi:10.1093/bioinformatics/btu827.
- Hershberger, L G, E G Shipley, and R K Meyer. 1953. "Myotrophic Activity of 19-Nortestosterone and Other Steroids Determined by Modified Levator Ani Muscle Method." *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine (New York, N.Y.)* 83 (1): 175–80. <http://www.ncbi.nlm.nih.gov/pubmed/13064212>.
- Heydari, Hamed, Naresh V R Mutha, Mahafizul Imran Mahmud, Cheuk Chuen Siow, Wei Yee Wee, Guat Jah Wong, Amir Hessam Yazdi, Mia Yang Ang, and Siew Woh Choo. 2014. "StaphyloBase: A Specialized Genomic Resource for the Staphylococcal Research Community." *Database : The Journal of Biological Databases and Curation* 2014 (January): bau010. doi:10.1093/database/bau010.
- Heydari, Hamed, Wei Yee Wee, Naline Lokanathan, Ranjeev Hari, Aini Mohamed Yusoff, Ching Yew Beh, Amir Hessam Yazdi, Guat Jah Wong, Yun Fong Ngeow, and Siew Woh Choo. 2013. "MabsBase: A Mycobacterium Abscessus Genome and Annotation Database." *PloS One* 8 (4): e62443. doi:10.1371/journal.pone.0062443.
- Hiller-Sturmhöfel, S, and A Bartke. 1998. "The Endocrine System: An Overview." *Alcohol Health and Research World* 22 (3): 153–64. <http://www.ncbi.nlm.nih.gov/pubmed/15706790>.
- Hossard, Laure, Laurence Guichard, Céline Pelosi, and David Makowski. 2017. "Lack of Evidence for a Decrease in Synthetic Pesticide Use on the Main Arable Crops in France." *Science of The Total Environment* 575 (January): 152–61. doi:10.1016/j.scitotenv.2016.10.008.
- Huber, Wolfgang, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, et al. 2015. "Orchestrating High-Throughput Genomic Analysis with Bioconductor." *Nature Methods* 12 (2): 115–21. doi:10.1038/nmeth.3252.
- Hull, K L, and S Harvey. 2000. "Growth Hormone: A Reproductive Endocrine-Paracrine Regulator?" *Reviews of Reproduction* 5 (3): 175–82. <http://www.ncbi.nlm.nih.gov/pubmed/11006167>.
- Hung, Hayley, Athanasios A Katsoyiannis, and Ramon Guardans. 2016. "Ten Years of Global Monitoring under the Stockholm Convention on Persistent Organic Pollutants (POPs): Trends, Sources and Transport Modelling." *Environmental Pollution (Barking, Essex : 1987)* 217 (October): 1–3. doi:10.1016/j.envpol.2016.05.035.
- Hutson, John M., Ruili Li, Bridget R. Southwell, Don Newgreen, and Mary Cousinery. 2015. "Regulation of Testicular Descent." *Pediatric Surgery International* 31 (4): 317–25. doi:10.1007/s00383-015-3673-4.

- Huyghe, Eric, Pierre Plante, and Patrick F Thonneau. 2007. "Testicular Cancer Variations in Time and Space in Europe." *European Urology* 51 (3): 621–28. doi:10.1016/j.eururo.2006.08.024.
- Igarashi, Yoshinobu, Noriyuki Nakatsu, Tomoya Yamashita, Atsushi Ono, Yasuo Ohno, Tetsuro Urushidani, and Hiroshi Yamada. 2015. "Open TG-GATES: A Large-Scale Toxicogenomics Database." *Nucleic Acids Research* 43 (Database issue): D921–7. doi:10.1093/nar/gku955.
- Irizarry, Rafael A, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. 2003. "Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data." *Biostatistics (Oxford, England)* 4 (2): 249–64. doi:10.1093/biostatistics/4.2.249.
- Ismail-Beigi, Faramarz, Patrick M Catalano, and Richard W Hanson. 2006. "Metabolic Programming: Fetal Origins of Obesity and Metabolic Syndrome in the Adult." *American Journal of Physiology. Endocrinology and Metabolism* 291 (3): E439–40. doi:10.1152/ajpendo.00105.2006.
- Ivell, R., and R. Anand-Ivell. 2009. "Biology of Insulin-like Factor 3 in Human Reproduction." *Human Reproduction Update* 15 (4): 463–76. doi:10.1093/humupd/dmp011.
- James, W. H. 2004. "Further Evidence That Mammalian Sex Ratios at Birth Are Partially Controlled by Parental Hormone Levels around the Time of Conception." *Human Reproduction* 19 (6): 1250–56. doi:10.1093/humrep/deh245.
- James, W H. 1995. "Offspring Sex Ratio as an Indicator of Reproductive Hazards Associated with Pesticides." *Occupational and Environmental Medicine* 52 (6): 429–30. <http://www.ncbi.nlm.nih.gov/pubmed/7627322>.
- Jégou, B. 1995. "La Cellule de Sertoli: Actualisation Du Concept de Cellule Nourricière." *Médecine/sciences* 11 (4): 519. doi:10.4267/10608/2241.
- Jégou, Bernard. 2015. "Paracetamol-Induced Endocrine Disruption in Human Fetal Testes." *Nature Reviews. Endocrinology* 11 (8): 453–54. doi:10.1038/nrendo.2015.106.
- Jensen, Morten S?ndergaard, Cristina Rebordosa, Ane Marie Thulstrup, Gunnar Toft, Henrik Toft S?rensen, Jens Peter Bonde, Tine Brink Henriksen, and J?rn Olsen. 2010. "Maternal Use of Acetaminophen, Ibuprofen, and Acetylsalicylic Acid During Pregnancy and Risk of Cryptorchidism." *Epidemiology* 21 (6): 779–85. doi:10.1097/EDE.0b013e3181f20bed.
- Jiang, Xuliang, James A. Dias, and Xiaolin He. 2014. "Structural Biology of Glycoprotein Hormones and Their Receptors: Insights to Signaling." *Molecular and Cellular Endocrinology* 382 (1): 424–51. doi:10.1016/j.mce.2013.08.021.
- Jin, Yuanxiang, Linggang Wang, Guanliang Chen, Xiaojian Lin, Wenyu Miao, and Zhengwei Fu. 2014. "Exposure of Mice to Atrazine and Its Metabolite Diaminochlorotriazine Elicits Oxidative Stress and Endocrine Disruption." *Environmental Toxicology and Pharmacology* 37 (2): 782–90. doi:10.1016/j.etap.2014.02.014.
- Jo, Sunhwan, Xi Cheng, Jumin Lee, Seonghoon Kim, Sang-Jun Park, Dhilon S. Patel, Andrew H. Beaven, et al. 2017. "CHARMM-GUI 10 Years for Biomolecular Modeling and Simulation." *Journal of Computational Chemistry* 38 (15): 1114–24. doi:10.1002/jcc.24660.
- Jobling, S, T Reynolds, R White, M G Parker, and J P Sumpter. 1995. "A Variety of Environmentally Persistent Chemicals, Including Some Phthalate Plasticizers, Are Weakly Estrogenic." *Environmental Health Perspectives* 103 (6): 582–87.

- <http://www.ncbi.nlm.nih.gov/pubmed/7556011>.
- Jones, K C, and P de Voogt. 1999. "Persistent Organic Pollutants (POPs): State of the Science." *Environmental Pollution (Barking, Essex : 1987)* 100 (1–3): 209–21.
<http://www.ncbi.nlm.nih.gov/pubmed/15093119>.
- Jones, R C, and R A Edgren. 1973. "The Effects of Various Steroids on the Vaginal Histology in the Rat." *Fertility and Sterility* 24 (4): 284–91.
<http://www.ncbi.nlm.nih.gov/pubmed/4694503>.
- Jönsson, Bo A G, Jonas Richthoff, Lars Rylander, Aleksander Giwercman, and Lars Hagmar. 2005. "Urinary Phthalate Metabolites and Biomarkers of Reproductive Function in Young Men." *Epidemiology (Cambridge, Mass.)* 16 (4): 487–93.
<http://www.ncbi.nlm.nih.gov/pubmed/15951666>.
- Kah, Olivier. 2016. *Les Perturbateurs Endocriniens : Ces Produits Qui En Veulent À Nos Hormones*. Éditions Apogée. <http://www.editions-apogee.com/perturbateurs-endocriniens-les.html>.
- Kang, Hwan Goo, Sang Hee Jeong, Joon Hyoung Cho, Dong Gyu Kim, Jong Myung Park, and Myung Haing Cho. 2005. "Evaluation of Estrogenic and Androgenic Activity of Butylated Hydroxyanisole in Immature Female and Castrated Rats." *Toxicology* 213 (1–2): 147–56. doi:10.1016/j.tox.2005.05.027.
- Kanno, J, L Onyon, J Haseman, P Fenner-Crisp, J Ashby, W Owens, and Organisation for Economic Co-operation and Development. 2001. "The OECD Program to Validate the Rat Uterotrophic Bioassay to Screen Compounds for in Vivo Estrogenic Responses: Phase 1." *Environmental Health Perspectives* 109 (8): 785–94.
<http://www.ncbi.nlm.nih.gov/pubmed/11564613>.
- Kanno, Jun, Lesley Onyon, Shyamal Peddada, John Ashby, Elard Jacob, and William Owens. 2003a. "The OECD Program to Validate the Rat Uterotrophic Bioassay. Phase 2: Coded Single-Dose Studies." *Environmental Health Perspectives* 111 (12): 1550–58.
<http://www.ncbi.nlm.nih.gov/pubmed/12948897>.
- . 2003b. "The OECD Program to Validate the Rat Uterotrophic Bioassay. Phase 2: Dose-Response Studies." *Environmental Health Perspectives* 111 (12): 1530–49.
<http://www.ncbi.nlm.nih.gov/pubmed/12948896>.
- Katritzky, Alan R, Dan C Fara, Ruslan O Petrukhin, Douglas B Tatham, Uko Maran, Andre Lomaka, and Mati Karelson. 2002. "The Present Utility and Future Potential for Medicinal Chemistry of QSAR/QSPR with Whole Molecule Descriptors." *Current Topics in Medicinal Chemistry* 2 (12): 1333–56.
<http://www.ncbi.nlm.nih.gov/pubmed/12470284>.
- Kavlock, R. J., G. Ankley, J. Blancato, M. Breen, R. Conolly, D. Dix, K. Houck, et al. 2007. "Computational Toxicology--A State of the Science Mini Review." *Toxicological Sciences* 103 (1): 14–27. doi:10.1093/toxsci/kfm297.
- Kavlock, R J, G P Daston, C DeRosa, P Fenner-Crisp, L E Gray, S Kaattari, G Lucier, et al. 1996. "Research Needs for the Risk Assessment of Health and Environmental Effects of Endocrine Disruptors: A Report of the U.S. EPA-Sponsored Workshop." *Environmental Health Perspectives*, August, 715–40. <http://www.ncbi.nlm.nih.gov/pubmed/8880000>.
- Kavlock, Robert J, David J Dix, Keith A Houck, Richard S Judson, Matt T Martin, and Ann M Richard. 2007. "ToxCast TM : Developing Predictive Signatures for Chemical Toxicity." <http://altweb.jhsph.edu/wc6/paper623.pdf>.
- Kawamura, Kazuhiro, Jin Kumagai, Satoko Sudo, Sang-Young Chun, Margareta Pisarska,

- Hiroki Morita, Jorma Toppari, et al. 2004. "Paracrine Regulation of Mammalian Oocyte Maturation and Male Germ Cell Survival." *Proceedings of the National Academy of Sciences of the United States of America* 101 (19): 7323–28. doi:10.1073/pnas.0307061101.
- Kier, L B, and L H Hall. 1981. "Derivation and Significance of Valence Molecular Connectivity." *Journal of Pharmaceutical Sciences* 70 (6): 583–89. <http://www.ncbi.nlm.nih.gov/pubmed/7252795>.
- Kim, Sunghwan, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, et al. 2015. "PubChem Substance and Compound Databases." *Nucleic Acids Research*, September. doi:10.1093/nar/gkv951.
- Klip, Helen, Janneke Verloop, Jan D van Gool, Marlies ETA Koster, Curt W Burger, Flora E van Leeuwen, and OMEGA Project Group. 2002. "Hypospadias in Sons of Women Exposed to Diethylstilbestrol in Utero: A Cohort Study." *The Lancet* 359 (9312): 1102–7. doi:10.1016/S0140-6736(02)08152-7.
- Knez, Jure, Roman Kranvogel, Barbara Pregl Breznik, Ernest Vončina, and Veljko Vlaisavljević. 2014. "Are Urinary Bisphenol A Levels in Men Related to Semen Quality and Embryo Development after Medically Assisted Reproduction?" *Fertility and Sterility* 101 (1): 215–221.e5. doi:10.1016/j.fertnstert.2013.09.030.
- Kohonen, Pekka, Emilio Benfenati, David Bower, Rebecca Ceder, Michael Crump, Kevin Cross, Roland C. Grafström, et al. 2013. "The ToxBank Data Warehouse: Supporting the Replacement of In Vivo Repeated Dose Systemic Toxicity Testing." *Molecular Informatics* 32 (1): 47–63. doi:10.1002/minf.201200114.
- Kolesnikov, Nikolay, Emma Hastings, Maria Keays, Olga Melnichuk, Y Amy Tang, Eleanor Williams, Mirosław Dylag, et al. 2015. "ArrayExpress Update--Simplifying Data Submissions." *Nucleic Acids Research* 43 (Database issue): D1113-6. doi:10.1093/nar/gku1057.
- Kretser, D M de, K L Loveland, T Meehan, M K O'Bryan, D J Phillips, and N G Wreford. 2001. "Inhibins, Activins and Follistatin: Actions on the Testis." *Molecular and Cellular Endocrinology* 180 (1–2): 87–92. <http://www.ncbi.nlm.nih.gov/pubmed/11451576>.
- Kringelum, Jens, Sonny Kim Kjaerulff, Søren Brunak, Ole Lund, Tudor I. Oprea, and Olivier Taboureau. 2016. "ChemProt-3.0: A Global Chemical Biology Diseases Mapping." *Database* 2016 (February): bav123. doi:10.1093/database/bav123.
- Kristensen, D M, L Lesné, V Le Fol, C Desdoits-Lethimonier, N Dejucq-Rainsford, H Leffers, and B Jégou. 2012. "Paracetamol (Acetaminophen), Aspirin (Acetylsalicylic Acid) and Indomethacin Are Anti-Androgenic in the Rat Foetal Testis." *International Journal of Andrology* 35 (3): 377–84. doi:10.1111/j.1365-2605.2012.01282.x.
- Kuhn, Robert M, David Haussler, and W James Kent. 2013. "The UCSC Genome Browser and Associated Tools." *Briefings in Bioinformatics* 14 (2): 144–61. doi:10.1093/bib/bbs038.
- Kvetnansky, Richard, Esther L Sabban, and Miklos Palkovits. 2009. "Catecholaminergic Systems in Stress: Structural and Molecular Genetic Approaches." *Physiological Reviews* 89 (2): 535–606. doi:10.1152/physrev.00042.2006.
- Laan, Mark J. van der, and Katherine S. Pollard. 2003. "A New Algorithm for Hybrid Hierarchical Clustering with Visualization and the Bootstrap." *Journal of Statistical Planning and Inference* 117: 275–303. doi:10.1016/S0378-3758(02)00388-9.
- Labrie, Fernand. 2004. "Adrenal Androgens and Intracrinology." *Seminars in Reproductive*

- Medicine* 22 (4): 299–309. doi:10.1055/s-2004-861547.
- Lamb, J., Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, et al. 2006. “The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease.” *Science* 313 (5795): 1929–35. doi:10.1126/science.1132939.
- Langfelder, Peter, Bin Zhang, and Steve Horvath. 2008. “Defining Clusters from a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R.” *Bioinformatics (Oxford, England)* 24 (5): 719–20. doi:10.1093/bioinformatics/btm563.
- Law, Vivian, Craig Knox, Yannick Djoumbou, Tim Jewison, An Chi Guo, Yifeng Liu, Adam Maciejewski, et al. 2014. “DrugBank 4.0: Shedding New Light on Drug Metabolism.” *Nucleic Acids Research* 42 (Database issue): D1091–7. doi:10.1093/nar/gkt1068.
- Lê, Sébastien, Julie Josse, and François Husson. 2008. “FactoMineR: An R Package for Multivariate Analysis.” *Journal of Statistical Software*. doi:10.18637/jss.v025.i01.
- Lea, Isabel A., Hui Gong, Anand Paleja, Asif Rashid, and Jennifer Fostel. 2017. “CEBS: A Comprehensive Annotated Database of Toxicological Data.” *Nucleic Acids Research* 45 (D1): D964–71. doi:10.1093/nar/gkw1077.
- Lee, Y-M, K-S Kim, D R Jacobs, and D-H Lee. 2017. “Persistent Organic Pollutants in Adipose Tissue Should Be Considered in Obesity Research.” *Obesity Reviews : An Official Journal of the International Association for the Study of Obesity* 18 (2): 129–39. doi:10.1111/obr.12481.
- Leeuwen, K. van, T.W. Schultz, T. Henry, B. Diderich, and G.D. Veith. 2009. “Using Chemical Categories to Fill Data Gaps in Hazard Assessment.” *SAR and QSAR in Environmental Research* 20 (3–4): 207–20. doi:10.1080/10629360902949179.
- Lepailleur, Alban, Guillaume Poezevara, and Ronan Bureau. 2013. “Automated Detection of Structural Alerts (Chemical Fragments) in (Eco)toxicology.” *Computational and Structural Biotechnology Journal* 5 (6): e201302013. doi:10.5936/csbj.201302013.
- Li, Heng-Hong, Daniel R. Hyde, Renxiang Chen, Pamela Heard, Carole L. Yauk, Jiri Aubrecht, and Albert J. Fornace. 2015. “Development of a Toxicogenomics Signature for Genotoxicity Using a Dose-Optimization and Informatics Strategy in Human Cells.” *Environmental and Molecular Mutagenesis* 56 (6): 505–19. doi:10.1002/em.21941.
- Lopez, Laureen M, David A Grimes, Mario Chen, Conrad Otterness, Carolyn Westhoff, Alison Edelman, and Frans M Helmerhorst. 2013. “Hormonal Contraceptives for Contraception in Overweight or Obese Women.” In *Cochrane Database of Systematic Reviews*, edited by Laureen M Lopez, CD008452. Chichester, UK: John Wiley & Sons, Ltd. doi:10.1002/14651858.CD008452.pub3.
- Luna-Acosta, A, H Budzinski, K Le Menach, H Thomas-Guyon, and P Bustamante. 2015. “Persistent Organic Pollutants in a Marine Bivalve on the Marennes-Oléron Bay and the Gironde Estuary (French Atlantic Coast) - Part 1: Bioaccumulation.” *The Science of the Total Environment* 514 (May): 500–510. doi:10.1016/j.scitotenv.2014.08.071.
- Lund, Lars, Malene C. Engebjerg, Lars Pedersen, Vera Ehrenstein, Mette Nørgaard, and Henrik Toft Sørensen. 2009. “Prevalence of Hypospadias in Danish Boys: A Longitudinal Study, 1977–2005.” *European Urology* 55 (5): 1022–26. doi:10.1016/j.eururo.2009.01.005.
- Maamar, Millissia Ben, Laurianne Lesné, Christèle Desdoits-Lethimonier, Isabelle Coiffec, Julie Lassurguère, Vincent Lavoué, Yoann Deceuninck, et al. 2015. “An Investigation of the Endocrine-Disruptive Effects of Bisphenol A in Human and Rat Fetal Testes.” Edited

- by Angel Nadal. *PLOS ONE* 10 (2): e0117226. doi:10.1371/journal.pone.0117226.
- Maamar, Millissia Ben, Laurianne Lesné, Kristin Hennig, Christèle Desdoits-Lethimonier, Karen R Kilcoyne, Isabelle Coiffec, Antoine D Rolland, et al. 2017. "Ibuprofen Results in Alterations of Human Fetal Testis Development." *Scientific Reports* 7 (March). Nature Publishing Group: 44184. doi:10.1038/srep44184.
- Madan, A. K., Sanjay Bajaj, and Harish Dureja. 2013. "Classification Models for Safe Drug Molecules." In *Methods in Molecular Biology (Clifton, N.J.)*, 930:99–124. doi:10.1007/978-1-62703-059-5_5.
- Maffini, Maricel V, Beverly S Rubin, Carlos Sonnenschein, and Ana M Soto. 2006. "Endocrine Disruptors and Reproductive Health: The Case of Bisphenol-A." *Molecular and Cellular Endocrinology* 254–255 (July): 179–86. doi:10.1016/j.mce.2006.04.033.
- Maglott, Donna, Jim Ostell, Kim D Pruitt, and Tatiana Tatusova. 2011. "Entrez Gene: Gene-Centered Information at NCBI." *Nucleic Acids Research* 39 (Database issue). Oxford University Press: D52-7. doi:10.1093/nar/gkq1237.
- Mahendroo, M S, and D W Russell. 1999. "Male and Female Isoenzymes of Steroid 5alpha-Reductase." *Reviews of Reproduction* 4 (3): 179–83. <http://www.ncbi.nlm.nih.gov/pubmed/10521155>.
- Main, K M, N E Skakkebaek, and J Toppari. 2009. "Cryptorchidism as Part of the Testicular Dysgenesis Syndrome: The Environmental Connection." *Endocrine Development* 14. Basel: KARGER: 167–73. doi:10.1159/000207485.
- Main, Katharina M, Gerda K Mortensen, Marko M Kaleva, Kirsten A Boisen, Ida N Damgaard, Marla Chellakooty, Ida M Schmidt, et al. 2006. "Human Breast Milk Contamination with Phthalates and Alterations of Endogenous Reproductive Hormones in Infants Three Months of Age." *Environmental Health Perspectives* 114 (2): 270–76. <http://www.ncbi.nlm.nih.gov/pubmed/16451866>.
- Mainwaring, WIP, SA Haining, and B Harper. 1988. "The Functions of Testosterone and Its Metabolites." *New Comprehensive*. <http://www.sciencedirect.com/science/article/pii/S0167730608606468>.
- Makanji, Yogeshwar, Craig A Harrison, and David M Robertson. 2011. "Feedback Regulation by Inhibins A and B of the Pituitary Secretion of Follicle-Stimulating Hormone." *Vitamins and Hormones* 85: 299–321. doi:10.1016/B978-0-12-385961-7.00014-7.
- Mallela, Ajay Raj, Rohini Koya, Shivashankara Kaniyoor Nagari, and Aswini Kumar Mohapatra. 2015. "Ethionamide: Unusual Cause of Hypothyroidism." *JOURNAL OF CLINICAL AND DIAGNOSTIC RESEARCH* 9 (8): OD08-9. doi:10.7860/JCDR/2015/13531.6331.
- Mandal, Prabir K. 2005. "Dioxin: A Review of Its Environmental Effects and Its Aryl Hydrocarbon Receptor Biology." *Journal of Comparative Physiology B* 175 (4): 221–30. doi:10.1007/s00360-005-0483-3.
- Manna, Pulak R., Matthew T. Dyson, Darrell W. Eubank, Barbara J. Clark, Enzo Lalli, Paolo Sassone-Corsi, Anthony J. Zeleznik, and Douglas M. Stocco. 2002. "Regulation of Steroidogenesis and the Steroidogenic Acute Regulatory Protein by a Member of the cAMP Response-Element Binding Protein Family." *Molecular Endocrinology* 16 (1): 184–99. doi:10.1210/mend.16.1.0759.
- Marchant, Carol A, Katharine A Briggs, and Anthony Long. 2008. "In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic." *Toxicology Mechanisms and Methods* 18 (2–3): 177–87.

- doi:10.1080/15376510701857320.
- Martin, Matthew T, Thomas B. Knudsen, David M. Reif, Keith A. Houck, Richard S. Judson, Robert J. Kavlock, and David J. Dix. 2011. "Predictive Model of Rat Reproductive Toxicity from ToxCast High Throughput Screening1." *Biology of Reproduction* 85 (2): 327–39. doi:10.1095/biolreprod.111.090977.
- Massart, F, and G Saggese. 2010. "Oestrogenic Mycotoxin Exposures and Precocious Pubertal Development." *International Journal of Andrology* 33 (2): 369–76. doi:10.1111/j.1365-2605.2009.01009.x.
- Mathias, Stephen L., Jarrett Hines-Kay, Jeremy J. Yang, Gergely Zahoransky-Kohalmi, Cristian G. Bologa, Oleg Ursu, and Tudor I. Oprea. 2013. "The CARLSBAD Database: A Confederated Database of Chemical Bioactivities." *Database: The Journal of Biological Databases and Curation* 2013. Oxford University Press. doi:10.1093/database/bat044.
- Mauduit, C., A. Florin, S. Amara, A. Bozec, B. Siddeek, S. Cunha, L. Meunier, et al. 2006. "Effets À Long Terme Des Perturbateurs Endocriniens Environnementaux Sur La Fertilité Masculine." *Gynécologie Obstétrique & Fertilité* 34 (10): 978–84. doi:10.1016/j.gyobfe.2006.08.010.
- Mazaud-Guittot, S?verine, Christophe Nicolas Nicolaz, Christ?le Desdoits-Lethimonier, Isabelle Coiffec, Millissia Ben Maamar, Patrick Balaguer, David M. Kristensen, et al. 2013. "Paracetamol, Aspirin, and Indomethacin Induce Endocrine Disturbances in the Human Fetal Testis Capable of Interfering With Testicular Descent." *The Journal of Clinical Endocrinology & Metabolism* 98 (11): E1757–67. doi:10.1210/jc.2013-2531.
- Mazaud-Guittot, Séverine, Christophe Nicolas Nicolaz, Christèle Desdoits-Lethimonier, Isabelle Coiffec, Millissia Ben Maamar, Patrick Balaguer, David M Kristensen, et al. 2013. "Paracetamol, Aspirin, and Indomethacin Induce Endocrine Disturbances in the Human Fetal Testis Capable of Interfering with Testicular Descent." *The Journal of Clinical Endocrinology and Metabolism* 98 (11): E1757-67. doi:10.1210/jc.2013-2531.
- McKinlay, R., J.A. A Plant, J.N.B. N B Bell, and N. Voulvoulis. 2008. "Endocrine Disrupting Pesticides: Implications for Risk Assessment." *Environment International* 34 (2): 168–83. doi:10.1016/j.envint.2007.07.013.
- McLachlan, JA. 1980. "Estrogens in the Environment." *ELSEVIER NORTH-HOLLAND, INC., NY 1980*.
<https://scholar.google.fr/scholar?cluster=10746118296793724261&hl=fr&oi=scholar&sa=X&ved=0ahUKEwj1htW8g6TTAhWEvRoKHWrBBd8QgAMIJCgCMAA>.
- Mediavilla, M D, E J Sanchez-Barcelo, D X Tan, L Manchester, and R J Reiter. 2010. "Basic Mechanisms Involved in the Anti-Cancer Effects of Melatonin." *Current Medicinal Chemistry* 17 (36): 4462–81. <http://www.ncbi.nlm.nih.gov/pubmed/21062257>.
- Meeker, John D., Shelley Ehrlich, Thomas L. Toth, Diane L. Wright, Antonia M. Calafat, Ana T. Trisini, Xiaoyun Ye, and Russ Hauser. 2010. "Semen Quality and Sperm DNA Damage in Relation to Urinary Bisphenol A among Men from an Infertility Clinic???" *Reproductive Toxicology* 30 (4): 532–39. doi:10.1016/j.reprotox.2010.07.005.
- Merelli, Ivan, Horacio Pérez-Sánchez, Sandra Gesing, and Daniele D'Agostino. 2014. "High-Performance Computing and Big Data in Omics-Based Medicine." *BioMed Research International* 2014 (January): 825649. doi:10.1155/2014/825649.
- Mesnage, R., N. Defarge, J. Spiroux de Vendômois, and G.E. Séralini. 2015. "Potential Toxic Effects of Glyphosate and Its Commercial Formulations below Regulatory Limits." *Food*

- and *Chemical Toxicology* 84 (October): 133–53. doi:10.1016/j.fct.2015.08.012.
- Miller, G. W. 2014. “Improving Reproducibility in Toxicology.” *Toxicological Sciences* 139 (1): 1–3. doi:10.1093/toxsci/kfu050.
- Mimeault, C., A.J. Woodhouse, X.-S. Miao, C.D. Metcalfe, T.W. Moon, and V.L. Trudeau. 2005. “The Human Lipid Regulator, Gemfibrozil Bioconcentrates and Reduces Testosterone in the Goldfish, *Carassius Auratus*.” *Aquatic Toxicology* 73 (1): 44–54. doi:10.1016/j.aquatox.2005.01.009.
- Miossec, P, F Archambeaud-Mouveroux, and M P Teissier. 1997. “[Inhibition of Steroidogenesis by Ketoconazole. Therapeutic Uses].” *Annales D’endocrinologie* 58 (6): 494–502. <http://www.ncbi.nlm.nih.gov/pubmed/9686009>.
- Miyagawa, Shinichi, Masaru Sato, and Taisen Iguchi. 2011. “Molecular Mechanisms of Induction of Persistent Changes by Estrogenic Chemicals on Female Reproductive Tracts and External Genitalia.” *The Journal of Steroid Biochemistry and Molecular Biology* 127 (1–2): 51–57. doi:10.1016/j.jsbmb.2011.03.009.
- Mocarelli, Paolo, Pier Mario Gerthoux, Donald G Patterson, Silvano Milani, Giuseppe Limonta, Maria Bertona, Stefano Signorini, et al. 2008. “Dioxin Exposure, from Infancy through Puberty, Produces Endocrine Disruption and Affects Human Semen Quality.” *Environmental Health Perspectives* 116 (1). National Institute of Environmental Health Science: 70–77. doi:10.1289/ehp.10399.
- Mombelli, E., and J. Devillers. 2010. “Evaluation of the OECD (Q)SAR Application Toolbox and Toxtree for Predicting and Profiling the Carcinogenic Potential of Chemicals.” *SAR and QSAR in Environmental Research* 21 (7–8): 731–52. doi:10.1080/1062936X.2010.528598.
- Moon, Hyun-Ju, Tae Seok Kang, Tae Sung Kim, Il Hyun Kang, Seung Hee Kim, and Soon Young Han. 2010. “OECD Validation of Phase-3 Hershberger Assay Using the Stimulated Weanling Male Rat in Korea.” *Journal of Applied Toxicology : JAT* 30 (4): 361–68. doi:10.1002/jat.1506.
- Mruk, Dolores D., and C. Yan Cheng. 2004. “Sertoli-Sertoli and Sertoli-Germ Cell Interactions and Their Significance in Germ Cell Movement in the Seminiferous Epithelium during Spermatogenesis.” *Endocrine Reviews* 25 (5): 747–806. doi:10.1210/er.2003-0022.
- Mulas, Francesca, Amy Li, David H. Sherr, and Stefano Monti. 2017. “Network-Based Analysis of Transcriptional Profiles from Chemical Perturbations Experiments.” *BMC Bioinformatics* 18 (S5): 130. doi:10.1186/s12859-017-1536-9.
- Mullard, Asher. 2017. “2016 FDA Drug Approvals.” *Nature Reviews Drug Discovery* 16 (2): 73–76. doi:10.1038/nrd.2017.14.
- Multigner, Luc, Philippe Kadhel, Florence Rouget, Pascal Blanchet, and Sylvaine Cordier. 2016. “Chlordecone Exposure and Adverse Effects in French West Indies Populations.” *Environmental Science and Pollution Research* 23 (1): 3–8. doi:10.1007/s11356-015-4621-5.
- Multigner, Luc, Jean Rodrigue Ndong, Arnaud Giusti, Marc Romana, Helene Delacroix-Maillard, Sylvaine Cordier, Bernard Jégou, Jean Pierre Thome, and Pascal Blanchet. 2010. “Chlordecone Exposure and Risk of Prostate Cancer.” *Journal of Clinical Oncology* 28 (21): 3457–62. doi:10.1200/JCO.2009.27.2153.
- “Must Try Harder.” 2012. *Nature* 483 (7391): 509–509. doi:10.1038/483509a.
- Myshkin, Eugene, Richard Brennan, Tatiana Khasanova, Tatiana Sitnik, Tatiana

- Serebriyskaya, Elena Litvinova, Alexey Guryanov, Yuri Nikolsky, Tatiana Nikolskaya, and Svetlana Bureeva. 2012. "Prediction of Organ Toxicity Endpoints by QSAR Modeling Based on Precise Chemical-Histopathology Annotations." *Chemical Biology & Drug Design* 80 (3). Blackwell Publishing Ltd: 406–16. doi:10.1111/j.1747-0285.2012.01411.x.
- Nagata, K, N Murayama, M Miyata, M Shimada, A Urahashi, Y Yamazoe, and R Kato. 1996. "Isolation and Characterization of a New Rat P450 (CYP3A18) cDNA Encoding P450(6)beta-2 Catalyzing Testosterone 6 Beta- and 16 Alpha-Hydroxylations." *Pharmacogenetics* 6 (1): 103–11. <http://www.ncbi.nlm.nih.gov/pubmed/8845857>.
- Nair, Anroop B, and Shery Jacob. 2016. "A Simple Practice Guide for Dose Conversion between Animals and Human." *Journal of Basic and Clinical Pharmacy* 7 (2). Medknow Publications: 27–31. doi:10.4103/0976-0105.177703.
- Nassouri, A.S., F. Archambeaud, and R. Desailoud. 2012. "Perturbateurs Endocriniens : Échos Des Congrès d'Endocrinologie 2012." *Annales d'Endocrinologie* 73 (October): S36–44. doi:10.1016/S0003-4266(12)70013-6.
- NCBI Resource Coordinators. 2017. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 45 (D1): D12–17. doi:10.1093/nar/gkw1071.
- NCBI Resource Coordinators. 2015. "Database Resources of the National Center for Biotechnology Information." *Nucleic Acids Research* 44 (D1): D7–19. doi:10.1093/nar/gkv1290.
- . 2016. "————." *Nucleic Acids Research* 44 (D1): D7–19. doi:10.1093/nar/gkv1290.
- Nelson, Lawrence M. 2009. "Primary Ovarian Insufficiency." *New England Journal of Medicine* 360 (6): 606–14. doi:10.1056/NEJMc0808697.
- Neph, Shane, M. Scott Kuehn, Alex P. Reynolds, Eric Haugen, Robert E. Thurman, Audra K. Johnson, Eric Rynes, et al. 2012. "BEDOPS: High-Performance Genomic Feature Operations." *Bioinformatics* 28 (14): 1919–20. doi:10.1093/bioinformatics/bts277.
- Noble, William S. 2006. "What Is a Support Vector Machine?" *Nature Biotechnology* 24 (12): 1565–67. doi:10.1038/nbt1206-1565.
- Norgil Damgaard, Ida, Katharina Maria Main, Jorma Toppari, and Niels E. Skakkeby. 2002. "Impact of Exposure to Endocrine Disruptors Inutero and in Childhood on Adult Reproduction." *Best Practice & Research Clinical Endocrinology & Metabolism* 16 (2): 289–309. doi:10.1053/beem.2002.0205.
- Nurmio, Mirja, Jenny Kallio, Marion Adam, Artur Mayerhofer, Jorma Toppari, and Kirsi Jahnukainen. 2012. "Peritubular Myoid Cells Have a Role in Postnatal Testicular Growth." *Spermatogenesis* 2 (2). Taylor & Francis: 79–87. doi:10.4161/spmg.20067.
- Nussey, Stephen., Saffron A. Whitehead, National Institutes of Health (U.S.). PubMed Central., and National Center for Biotechnology Information (U.S.). 2001. *Endocrinology : An Integrated Approach*. Bios.
- O'Donnell, L, R I McLachlan, N G Wreford, D M de Kretser, and D M Robertson. 1996. "Testosterone Withdrawal Promotes Stage-Specific Detachment of Round Spermatids from the Rat Seminiferous Epithelium." *Biology of Reproduction* 55 (4): 895–901. <http://www.ncbi.nlm.nih.gov/pubmed/8879506>.
- Oatley, Jon M, and Ralph L Brinster. 2012. "The Germline Stem Cell Niche Unit in Mammalian Testes." *Physiological Reviews* 92 (2). NIH Public Access: 577–95. doi:10.1152/physrev.00025.2011.

- OECD. 2012. "OECD Conceptual Framework for Testing and Assessment of Endocrine Disrupters (as Revised in 2012)." *OECD Environmental Health and Safety Publications Series on Testing and Assessment*, no. 150. [https://www.oecd.org/env/ehs/testing/OECD Conceptual Framework for Testing and Assessment of Endocrine Disrupters for the public website.pdf](https://www.oecd.org/env/ehs/testing/OECD_Conceptual_Framework_for_Testing_and_Assessment_of_Endocrine_Disrupters_for_the_public_website.pdf).
- Orvis, Joshua, Jonathan Crabtree, Kevin Galens, Aaron Gussman, Jason M Inman, Eduardo Lee, Sreenath Nampally, et al. 2010. "Ergatis: A Web Interface and Scalable Software System for Bioinformatics Workflows." *Bioinformatics (Oxford, England)* 26 (12): 1488–92. doi:10.1093/bioinformatics/btq167.
- Oury, Franck, Grzegorz Sumara, Olga Sumara, Mathieu Ferron, Haixin Chang, Charles E Smith, Louis Hermo, et al. 2011. "Endocrine Regulation of Male Fertility by the Skeleton." *Cell* 144 (5). NIH Public Access: 796–809. doi:10.1016/j.cell.2011.02.004.
- Owens, William, L Earl Gray, Errol Zeiger, Michael Walker, Kanji Yamasaki, John Ashby, and Elard Jacob. 2007. "The OECD Program to Validate the Rat Hershberger Bioassay to Screen Compounds for in Vivo Androgen and Antiandrogen Responses: Phase 2 Dose-Response Studies." *Environmental Health Perspectives* 115 (5): 671–78. doi:10.1289/ehp.9666.
- Palmer, J. R., Lauren A Wise, Elizabeth E Hatch, Rebecca Troisi, Linda Titus-Ernstoff, William Strohsnitter, Raymond Kaufman, et al. 2006. "Prenatal Diethylstilbestrol Exposure and Risk of Breast Cancer." *Cancer Epidemiology Biomarkers & Prevention* 15 (8): 1509–14. doi:10.1158/1055-9965.EPI-06-0109.
- Pamies, David, and Thomas Hartung. 2017. "21st Century Cell Culture for 21st Century Toxicology." *Chemical Research in Toxicology* 30 (1): 43–52. doi:10.1021/acs.chemrestox.6b00269.
- Parada, Luis F., and Serge Nef. 1999. "Cryptorchidism in Mice Mutant for Insl3." *Nature Genetics* 22 (3): 295–99. doi:10.1038/10364.
- Parks, L G, J S Ostby, C R Lambright, B D Abbott, G R Klinefelter, N J Barlow, and L E Gray. 2000. "The Plasticizer Diethylhexyl Phthalate Induces Malformations by Decreasing Fetal Testosterone Synthesis during Sexual Differentiation in the Male Rat." *Toxicological Sciences : An Official Journal of the Society of Toxicology* 58 (2): 339–49. <http://www.ncbi.nlm.nih.gov/pubmed/11099646>.
- Patlewicz, G, N Jeliaskova, R J Safford, A P Worth, and B Aleksiev. 2008. "An Evaluation of the Implementation of the Cramer Classification Scheme in the Toxtree Software." *SAR and QSAR in Environmental Research* 19 (5–6): 495–524. doi:10.1080/10629360802083871.
- Paula, F J A de, and C J Rosen. 2010. "Back to the Future: Revisiting Parathyroid Hormone and Calcitonin Control of Bone Remodeling." *Hormone and Metabolic Research = Hormon- Und Stoffwechselforschung = Hormones et Metabolisme* 42 (5): 299–306. doi:10.1055/s-0030-1248255.
- Payne, A H, and G L Youngblood. 1995. "Regulation of Expression of Steroidogenic Enzymes in Leydig Cells." *Biology of Reproduction* 52 (2): 217–25. <http://www.ncbi.nlm.nih.gov/pubmed/7711191>.
- Peart, W S. 1977. "The Kidney as an Endocrine Organ." *Lancet (London, England)* 2 (8037): 543–48. <http://www.ncbi.nlm.nih.gov/pubmed/95741>.
- Pedersen, Finn, Jack De Bruijn, Sharon Munn, and Kees Van Leeuwen. 2003. "Assessment of Additional Testing Needs under REACH Effects of (Q)SARS, Risk Based Testing and

- Voluntary Industry Initiatives.”
<http://home.kpn.nl/reach/downloads/reachtestingneedsfinal.pdf>.
- Pelletier, G, C Labrie, and F Labrie. 2000. “Localization of Oestrogen Receptor Alpha, Oestrogen Receptor Beta and Androgen Receptors in the Rat Reproductive Organs.” *The Journal of Endocrinology* 165 (2): 359–70.
<http://www.ncbi.nlm.nih.gov/pubmed/10810300>.
- Penman, Mike, Marcy Banton, Steffen Erler, Nigel Moore, and Klaus Semmler. 2015. “Olefins and Chemical Regulation in Europe: REACH.” *Chemico-Biological Interactions* 241 (April): 59–65. doi:10.1016/j.cbi.2015.04.001.
- Pérez, Luis Orlando, Rolando González-José, and Pilar Peral García. 2016. “Prediction of Non-Genotoxic Carcinogenicity Based on Genetic Profiles of Short Term Exposure Assays.” *Toxicological Research* 32 (4): 289–300. doi:10.5487/TR.2016.32.4.289.
- Pesatori, Angela Cecilia, Dario Consonni, Silvia Bachetti, Carlo Zocchetti, Matteo Bonzini, Andrea Baccarelli, and Pier Alberto Bertazzi. 2003. “Short- and Long-Term Morbidity and Mortality in the Population Exposed to Dioxin after the ‘Seveso Accident’.” *Industrial Health* 41 (3): 127–38.
<http://www.ncbi.nlm.nih.gov/pubmed/12916742>.
- Petersen, Cecilia, and Olle Soder. 2006. “The Sertoli Cell--a Hormonal Target and ‘Super’ Nurse for Germ Cells That Determines Testicular Size.” *Hormone Research* 66 (4): 153–61. doi:10.1159/000094142.
- Pfaller, Walter, Pilar Prieto, Wolfgang Dekant, Paul Jennings, and Bas J. Blaauboer. 2015. “The Predict-IV Project: Towards Predictive Toxicology Using in Vitro Techniques.” *Toxicology in Vitro* 30 (1): 1–3. doi:10.1016/j.tiv.2015.10.006.
- Pitetti, Jean-Luc, Pierre Calvel, Céline Zimmermann, Béatrice Conne, Marilena D. Papaioannou, Florence Aubry, Christopher R. Cederroth, et al. 2013. “An Essential Role for Insulin and IGF1 Receptors in Regulating Sertoli Cell Proliferation, Testis Size, and FSH Action in Mice.” *Molecular Endocrinology* 27 (5): 814–27. doi:10.1210/me.2012-1258.
- Pizzo, Fabiola, Anna Lombardo, Alberto Manganaro, and Emilio Benfenati. 2013. “In Silico Models for Predicting Ready Biodegradability under REACH: A Comparative Study.” *The Science of the Total Environment* 463–464 (October): 161–68.
doi:10.1016/j.scitotenv.2013.05.060.
- Plan Ecophyto. 2008. “Plan ECOPHYTO 2018 de Réduction de L’usage Des Pesticides 2008–2018.” http://www.observatoire-pesticides.gouv.fr/upload/bibliotheque/110786233307360592329934193591/PLAN_ECOPHYTO_2018.pdf.
- Plan Ecophyto II. 2015. “Plan Ecophyto II.” http://www.agence-nationale-recherche.fr/fileadmin/documents/2017/151022_ecophyto.pdf.
- Poland, Craig A, Mark R Miller, Rodger Duffin, and Flemming Cassee. 2014. “The Elephant in the Room: Reproducibility in Toxicology.” *Particle and Fibre Toxicology* 11 (1): 42. doi:10.1186/s12989-014-0042-8.
- Pop, Anca, Bela Kiss, and Felicia Loghin. 2013. “Endocrine Disrupting Effects of Butylated Hydroxyanisole (BHA - E320).” *Chujul Medical (1957)* 86 (1): 16–20.
<http://www.ncbi.nlm.nih.gov/pubmed/26527908>.
- Porter, M. P., M. K. Faizan, R. W. Grady, and B. A. Mueller. 2005. “Hypospadias in Washington State: Maternal Risk Factors and Prevalence Trends.” *PEDIATRICS* 115 (4):

- e495–99. doi:10.1542/peds.2004-1552.
- Portier, Christopher J, Bruce K Armstrong, Bruce C Baguley, Xaver Baur, Igor Belyaev, Robert Bellé, Fiorella Belpoggi, et al. 2016. “Differences in the Carcinogenic Evaluation of Glyphosate between the International Agency for Research on Cancer (IARC) and the European Food Safety Authority (EFSA).” *Journal of Epidemiology and Community Health* 70 (8): 741–45. doi:10.1136/jech-2015-207005.
- Prathipati, Philip, and Kenji Mizuguchi. 2016. “Systems Biology Approaches to a Rational Drug Discovery Paradigm.” *Current Topics in Medicinal Chemistry* 16 (9): 1009–25. <http://www.ncbi.nlm.nih.gov/pubmed/26306988>.
- Prlić, Andreas, and James B Procter. 2012. “Ten Simple Rules for the Open Development of Scientific Software.” *PLoS Computational Biology* 8 (12): e1002802. doi:10.1371/journal.pcbi.1002802.
- Prunet-Marcassus, Bénédicte, Mathieu Desbazeille, Arnaud Bros, Katie Louche, Philippe Delagrangé, Pierre Renard, Louis Casteilla, and Luc Pénicaud. 2003. “Melatonin Reduces Body Weight Gain in Sprague Dawley Rats with Diet-Induced Obesity.” *Endocrinology* 144 (12): 5347–52. doi:10.1210/en.2003-0693.
- Quinlan, Aaron R. 2014. “BEDTools: The Swiss-Army Tool for Genome Feature Analysis.” *Current Protocols in Bioinformatics / Editorial Board, Andreas D. Baxevanis ... [et Al.]* 47 (January): 11.12.1-11.12.34. doi:10.1002/0471250953.bi1112s47.
- R core Team. 2013. *R: A Language and Environment for Statistical Computing*. <http://www.gbif.org/resource/81287>.
- Randic, Milan. 1975. “On Characterization of Molecular Branching.” <http://pubs.acs.org/doi/pdf/10.1021/ja00856a001>.
- Ravenzwaay, Bennard van, Susanne N Kolle, Tzutzuy Ramirez, and Hennicke G Kamp. 2013. “Vinclozolin: A Case Study on the Identification of Endocrine Active Substances in the Past and a Future Perspective.” *Toxicology Letters* 223 (3): 271–79. doi:10.1016/j.toxlet.2013.03.029.
- Reiter, Russel J., Dun-Xian Tan, and Lorena Fuentes-Broto. 2010. “Melatonin: A Multitasking Molecule.” In *Progress in Brain Research*, 181:127–51. doi:10.1016/S0079-6123(08)81008-4.
- Richard, Ann M., Richard S. Judson, Keith A. Houck, Christopher M. Grulke, Patra Volarath, Inthirany Thillainadarajah, Chihai Yang, et al. 2016. “ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology.” *Chemical Research in Toxicology* 29 (8): 1225–51. doi:10.1021/acs.chemrestox.6b00135.
- Richards, Joanne S, and Stephanie A Pangas. 2010. “The Ovary: Basic Biology and Clinical Implications.” *The Journal of Clinical Investigation* 120 (4): 963–72. doi:10.1172/JCI41350.
- Ritchie, Matthew E, Belinda Phipson, Di Wu, Yifang Hu, Charity W Law, Wei Shi, and Gordon K Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47. doi:10.1093/nar/gkv007.
- Robertson, Courtney, Bruce D. Pauli, Vance L. Trudeau, and Laia Navarro-Martín. 2016. “Characterization and Developmental Expression Profile of the Steroidogenic Acute Regulatory Protein (StAR) in the Gonad-Mesonephros Complex of <i>Lithobates Sylvaticus</i>.” *Sexual Development* 10 (2): 91–96. doi:10.1159/000445816.

- Robinson, M K, R Osborne, and M A Perkins. 2000. "In Vitro and Human Testing Strategies for Skin Irritation." *Annals of the New York Academy of Sciences* 919: 192–204. <http://www.ncbi.nlm.nih.gov/pubmed/11083109>.
- Rogers, J M, and M S Denison. 2000. "Recombinant Cell Bioassays for Endocrine Disruptors: Development of a Stably Transfected Human Ovarian Cell Line for the Detection of Estrogenic and Anti-Estrogenic Chemicals." *In Vitro & Molecular Toxicology* 13 (1): 67–82. <http://www.ncbi.nlm.nih.gov/pubmed/10900408>.
- Romani, Federica, Anna Tropea, Elisa Scarinci, Cinzia Dello Russo, Lucia Lisi, Stefania Catino, Antonio Lanzone, and Rosanna Apa. 2013. "Endocrine Disruptors and Human Corpus Luteum: In Vitro Effects of Phenols on Luteal Cells Function." *Journal of Environmental Science and Health. Part C, Environmental Carcinogenesis & Ecotoxicology Reviews* 31 (2): 170–80. doi:10.1080/10590501.2013.782180.
- Römer, Michael, Linus Backert, Johannes Eichner, and Andreas Zell. 2014. "ToxDBScan: Large-Scale Similarity Screening of Toxicological Databases for Drug Candidates." *International Journal of Molecular Sciences* 15 (10): 19037–55. doi:10.3390/ijms151019037.
- Roncaglioni, Alessandra, Andrey A Toropov, Alla P Toropova, and Emilio Benfenati. 2013. "In Silico Methods to Predict Drug Toxicity." *Current Opinion in Pharmacology* 13 (5): 802–6. doi:10.1016/j.coph.2013.06.001.
- Rosenbloom, Kate R, Joel Armstrong, Galt P Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R Dreszer, et al. 2014. "The UCSC Genome Browser Database: 2015 Update." *Nucleic Acids Research* 43 (Database issue): D670–81. doi:10.1093/nar/gku1177.
- Rosol, Thomas J., John T. Yarrington, John Latendresse, and Charles C. Capen. 2001. "Adrenal Gland: Structure, Function, and Mechanisms of Toxicity." *Toxicologic Pathology* 29 (1): 41–48. doi:10.1080/019262301301418847.
- Rosse, Cornelius, and José L.V. Mejino. 2003. "A Reference Ontology for Biomedical Informatics: The Foundational Model of Anatomy." *Journal of Biomedical Informatics* 36 (6): 478–500. doi:10.1016/j.jbi.2003.11.007.
- Rother, Kristian, Wojciech Potrzebowski, Tomasz Puton, Magdalena Rother, Ewa Wywiał, and Janusz M Bujnicki. 2012. "A Toolbox for Developing Bioinformatics Software." *Briefings in Bioinformatics* 13 (2). Oxford University Press: 244–57. doi:10.1093/bib/bbr035.
- Rowland, M. 2013. "Physiologically-Based Pharmacokinetic (PBPK) Modeling and Simulations Principles, Methods, and Applications in the Pharmaceutical Industry." *CPT: Pharmacometrics & Systems Pharmacology* 2 (7): e55. doi:10.1038/psp.2013.29.
- Rubin, Beverly S. 2011. "Bisphenol A: An Endocrine Disruptor with Widespread Exposure and Multiple Effects." *The Journal of Steroid Biochemistry and Molecular Biology* 127 (1–2): 27–34. doi:10.1016/j.jsbmb.2011.05.002.
- Safe, S. 2001. "Molecular Biology of the Ah Receptor and Its Role in Carcinogenesis." *Toxicology Letters* 120 (1–3): 1–7. <http://www.ncbi.nlm.nih.gov/pubmed/11323156>.
- Safran, Marilyn, Irina Solomon, Orit Shmueli, Michal Lapidot, Shai Shen-Orr, Avital Adato, Uri Ben-Dor, et al. 2002. "GeneCards 2002: Towards a Complete, Object-Oriented, Human Gene Compendium." *Bioinformatics (Oxford, England)* 18 (11): 1542–43. <http://www.ncbi.nlm.nih.gov/pubmed/12424129>.
- Sagnella, Guiseppe A. 2002. "The Heart as an Endocrine Organ." *Biologist (London, England)*

- 49 (6): 275–79. <http://www.ncbi.nlm.nih.gov/pubmed/12486305>.
- Sarntivijai, Sirarat, Yu Lin, Zuoshuang Xiang, Terrence F Meehan, Alexander D Diehl, Uma D Vempati, Stephan C Schürer, et al. 2014. “CLO: The Cell Line Ontology.” *Journal of Biomedical Semantics* 5 (January): 37. doi:10.1186/2041-1480-5-37.
- Schettler, Ted. 2006. “Human Exposure to Phthalates via Consumer Products.” *International Journal of Andrology* 29 (1): 134–39. doi:10.1111/j.1365-2605.2005.00567.x.
- Seki, Masanori, Hirofumi Yokota, Haruki Matsubara, Masanobu Maeda, Hiroshi Tadokoro, and Kunio Kobayashi. 2004. “Fish Full Life-Cycle Testing for Androgen Methyltestosterone on Medaka (*Oryzias Latipes*).” *Environmental Toxicology and Chemistry* 23 (3): 774–81. <http://www.ncbi.nlm.nih.gov/pubmed/15285372>.
- Séralini, Gilles-Eric, Emilie Clair, Robin Mesnage, Steeve Gress, Nicolas Defarge, Manuela Malatesta, Didier Hennequin, and Joël Spiroux de Vendômois. 2012. “RETRACTED: Long Term Toxicity of a Roundup Herbicide and a Roundup-Tolerant Genetically Modified Maize.” *Food and Chemical Toxicology* 50 (11): 4221–31. doi:10.1016/j.fct.2012.08.005.
- Sheikh, Ishfaq A., Muhammad Yasir, Muhammad Abu-Elmagd, Tanveer A. Dar, Adel M. Abuzenadah, Ghazi A. Damanhour, Mohammed Al-Qahtani, and Mohd A. Beg. 2016. “Human Sex Hormone-Binding Globulin as a Potential Target of Alternate Plasticizers: An in Silico Study.” *BMC Structural Biology* 16 (S1): 15. doi:10.1186/s12900-016-0067-3.
- Shi, Wei, Dongyang Deng, Yuting Wang, Guanjiu Hu, Jing Guo, Xiaowei Zhang, Xinru Wang, John P. Giesy, Hongxia Yu, and Ziheng Wang. 2016. “Causes of Endocrine Disrupting Potencies in Surface Water in East China.” *Chemosphere* 144 (February): 1435–42. doi:10.1016/j.chemosphere.2015.09.018.
- Simpson, E R. 2000. “Role of Aromatase in Sex Steroid Action.” *Journal of Molecular Endocrinology* 25 (2): 149–56. <http://www.ncbi.nlm.nih.gov/pubmed/11013343>.
- Sinclair, Adriane Watkins, Mei Cao, Andrew Pask, Laurence Baskin, and Gerald R. Cunha. 2017. “Flutamide-Induced Hypospadias in Rats: A Critical Assessment.” *Differentiation* 94 (March): 37–57. doi:10.1016/j.diff.2016.12.001.
- Skakkebaek, N E, E Rajpert-De Meyts, and K M Main. 2001. “Testicular Dysgenesis Syndrome: An Increasingly Common Developmental Disorder with Environmental Aspects.” *Human Reproduction (Oxford, England)* 16 (5): 972–78. <http://www.ncbi.nlm.nih.gov/pubmed/11331648>.
- Skakkebaek, Niels E, Ewa Rajpert-De Meyts, Germaine M Buck Louis, Jorma Toppari, Anna-Maria Andersson, Michael L Eisenberg, Tina Kold Jensen, et al. 2016. “Male Reproductive Disorders and Fertility Trends: Influences of Environment and Genetic Susceptibility.” *Physiological Reviews* 96 (1). American Physiological Society: 55–97. doi:10.1152/physrev.00017.2015.
- Skinner, Mitchell E, Andrew V Uzilov, Lincoln D Stein, Christopher J Mungall, and Ian H Holmes. 2009. “JBrowse: A next-Generation Genome Browser.” *Genome Research* 19 (9): 1630–38. doi:10.1101/gr.094607.109.
- Smith, Barry, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J Goldberg, et al. 2007. “The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration.” *Nature Biotechnology* 25 (11): 1251–55. doi:10.1038/nbt1346.
- Snijder, Claudia A, Andreas Kortenkamp, Eric A P Steegers, Vincent W V Jaddoe, Albert

- Hofman, Ulla Hass, and Alex Burdorf. 2012. "Intrauterine Exposure to Mild Analgesics during Pregnancy and the Occurrence of Cryptorchidism and Hypospadias in the Offspring: The Generation R Study." *Human Reproduction (Oxford, England)* 27 (4): 1191–1201. doi:10.1093/humrep/der474.
- Sogani, P C, M R Vagaiwala, and W F Whitmore. 1984. "Experience with Flutamide in Patients with Advanced Prostatic Cancer without Prior Endocrine Therapy." *Cancer* 54 (4): 744–50. <http://www.ncbi.nlm.nih.gov/pubmed/6378356>.
- Solomon, Keith R, James A Carr, Louis H Du Preez, John P Giesy, Ronald J Kendall, Ernest E Smith, and Glen J Van Der Kraak. 2008. "Effects of Atrazine on Fish, Amphibians, and Aquatic Reptiles: A Critical Review." *Critical Reviews in Toxicology* 38 (9): 721–72. doi:10.1080/10408440802116496.
- Song, Mee, Youn-Jung Kim, Yong-Keun Park, and Jae-Chun Ryu. 2012. "Changes in Thyroid Peroxidase Activity in Response to Various Chemicals." *Journal of Environmental Monitoring* 14 (8): 2121. doi:10.1039/c2em30106g.
- Sonneveld, Edwin, Jacoba A C Riteco, Hendrina J Jansen, Bart Pieterse, Abraham Brouwer, Willem G Schoonen, and Bart van der Burg. 2006. "Comparison of in Vitro and in Vivo Screening Models for Androgenic and Estrogenic Activities." *Toxicological Sciences : An Official Journal of the Society of Toxicology* 89 (1): 173–87. doi:10.1093/toxsci/kfj009.
- Srinivasan, Venkatramanujam, Warren D. Spence, Seithikurippu R. Pandi-Perumal, Rahima Zakharia, Kunwar P. Bhatnagar, and Amnon Brzezinski. 2009. "Melatonin and Human Reproduction: Shedding Light on the Darkness Hormone." *Gynecological Endocrinology* 25 (12): 779–85. doi:10.3109/09513590903159649.
- Steiner, Guido, Laura Suter, Franziska Boess, Rodolfo Gasser, Maria Cristina de Vera, Silvio Albertini, and Stefan Ruepp. 2004. "Discriminating Different Classes of Toxicants by Transcript Profiling." *Environmental Health Perspectives*. August. <http://www.ncbi.nlm.nih.gov/pubmed/?term=Discriminating+different+classes+of+tox+icants+by+transcript+profiling>.
- Steinmetz, Fabian P., Judith C. Madden, and Mark T. D. Cronin. 2015. "Data Quality in the Human and Environmental Health Sciences: Using Statistical Confidence Scoring to Improve QSAR/QSPR Modeling." *Journal of Chemical Information and Modeling* 55 (8): 1739–46. doi:10.1021/acs.jcim.5b00294.
- Stocco, Douglas M. 2000. "StARTing to Understand Cholesterol Transfer." *Nature Structural Biology* 7 (6): 445–47. doi:10.1038/75834.
- Sumpter, J P, and S Jobling. 1995. "Vitellogenesis as a Biomarker for Estrogenic Contamination of the Aquatic Environment." *Environmental Health Perspectives*, October, 173–78. <http://www.ncbi.nlm.nih.gov/pubmed/8593867>.
- Sundell, David, Chanaka Mannapperuma, Sergiu Netotea, Nicolas Delhomme, Yao-Cheng Lin, Andreas Sjödin, Yves Van de Peer, Stefan Jansson, Torgeir R. Hvidsten, and Nathaniel R. Street. 2015. "The Plant Genome Integrative Explorer Resource: PlantGenIE.org." *New Phytologist* 208 (4): 1149–56. doi:10.1111/nph.13557.
- Suter, Laura, Lee E Babiss, and Eric B Wheeldon. 2004. "Toxicogenomics in Predictive Toxicology in Drug Development." *Chemistry & Biology* 11 (2): 161–71. doi:10.1016/j.chembiol.2004.02.003.
- Swan, Shanna H. 2008. "Environmental Phthalate Exposure in Relation to Reproductive Outcomes and Other Health Endpoints in Humans." *Environmental Research* 108 (2):

- 177–84. <http://www.ncbi.nlm.nih.gov/pubmed/18949837>.
- Taboureau, Olivier, Anne Hersey, Karine Audouze, Laurent Gautier, Ulrik P. Jacobsen, Ruth Akhtar, Francis Atkinson, John P. Overington, and Søren Brunak. 2012. “Toxicogenomics Investigation Under the eTOX Project.” *Journal of Pharmacogenomics & Pharmacoproteomics* 7 (January). OMICS International: 1–5. doi:10.4172/2153-0645.S7-001.
- Takeuchi, Toru, Osamu Tsutsumi, Yumiko Ikezuki, Yasushi Takai, and Yuji Taketani. 2004. “Positive Relationship between Androgen and the Endocrine Disruptor, Bisphenol A, in Normal Women and Women with Ovarian Dysfunction.” *Endocrine Journal* 51 (2): 165–69. <http://www.ncbi.nlm.nih.gov/pubmed/15118266>.
- Talmage, Roy V, and H T Mobley. 2008. “Calcium Homeostasis: Reassessment of the Actions of Parathyroid Hormone.” *General and Comparative Endocrinology* 156 (1): 1–8. doi:10.1016/j.ygcen.2007.11.003.
- Taylor, J R. 1985. “Neurological Manifestations in Humans Exposed to Chlordecone: Follow-up Results.” *Neurotoxicology* 6 (1): 231–36. <http://www.ncbi.nlm.nih.gov/pubmed/2581197>.
- Theil, Frank-Peter, Theodor W Guentert, Sami Haddad, and Patrick Poulin. 2003. “Utility of Physiologically Based Pharmacokinetic Models to Drug Development and Rational Drug Discovery Candidate Selection.” *Toxicology Letters* 138 (1–2): 29–49. <http://www.ncbi.nlm.nih.gov/pubmed/12559691>.
- Thompson, E W, A W Blackshaw, and S S Raychoudhury. 1995. “Secreted Products and Extracellular Matrix from Testicular Peritubular Myoid Cells Influence Androgen-Binding Protein Secretion by Sertoli Cells in Culture.” *Journal of Andrology* 16 (1): 28–35. <http://www.ncbi.nlm.nih.gov/pubmed/7768750>.
- Thompson, Matthew D, and Daniel A Beard. 2011. “Development of Appropriate Equations for Physiologically Based Pharmacokinetic Modeling of Permeability-Limited and Flow-Limited Transport.” *Journal of Pharmacokinetics and Pharmacodynamics* 38 (4). NIH Public Access: 405–21. doi:10.1007/s10928-011-9200-x.
- Thorvaldsdottir, H., J. T. Robinson, and J. P. Mesirov. 2013. “Integrative Genomics Viewer (IGV): High-Performance Genomics Data Visualization and Exploration.” *Briefings in Bioinformatics* 14 (2): 178–92. doi:10.1093/bib/bbs017.
- Timms, Barry G, Kembra L Howdeshell, Lesley Barton, Sarahann Bradley, Catherine A Richter, and Frederick S vom Saal. 2005. “Estrogenic Chemicals in Plastic and Oral Contraceptives Disrupt Development of the Fetal Mouse Prostate and Urethra.” *Proceedings of the National Academy of Sciences of the United States of America* 102 (19): 7014–19. doi:10.1073/pnas.0502544102.
- Toda, Katsumi, Teruhiko Okada, Chisata Miyaura, and Toshiji Saibara. 2003. “Fenofibrate, a Ligand for PPARalpha, Inhibits Aromatase Cytochrome P450 Expression in the Ovary of Mouse.” *Journal of Lipid Research* 44 (2): 265–70. doi:10.1194/jlr.M200327-JLR200.
- Todeschini, R, and V Consonni. 2008. “Handbook of Molecular Descriptors.” <https://books.google.fr/books?hl=fr&lr=&id=TCuHqbvvgMbEC&oi=fnd&pg=PP2&dq=Handbook+of+Molecular+Descriptors&ots=juHCxewNj9&sig=3irqICPUfV2ZWih76OYkkRLzhw>.
- Trasande, Leonardo, R. Thomas Zoeller, Ulla Hass, Andreas Kortenkamp, Philippe Grandjean, John Peterson Myers, Joseph DiGangi, et al. 2015. “Estimating Burden and Disease Costs of Exposure to Endocrine-Disrupting Chemicals in the European Union.”

- The Journal of Clinical Endocrinology & Metabolism* 100 (4): 1245–55.
doi:10.1210/jc.2014-4324.
- Tropsha, Alexander, Paola Gramatica, and Vijay K Gombar. 2003. “The Importance of Being Earnest: Validation Is the Absolute Essential for Successful Application and Interpretation of QSPR Models.” *QSAR & Combinatorial Science* 22 (1). WILEY-VCH Verlag: 69–77. doi:10.1002/qsar.200390007.
- Tullner, W W. 1961. “Uterotrophic Action of the Insecticide Methoxychlor.” *Science (New York, N.Y.)* 133 (3453): 647. <http://www.ncbi.nlm.nih.gov/pubmed/13778585>.
- Uhlen, M., L. Fagerberg, B. M. Hallstrom, C. Lindskog, P. Oksvold, A. Mardinoglu, A. Sivertsson, et al. 2015. “Tissue-Based Map of the Human Proteome.” *Science* 347 (6220): 1260419–1260419. doi:10.1126/science.1260419.
- Usepa. 1998. “Endocrine Disruptor Screening and Testing Advisory Committee Final Report - Table of Contents.” <https://www.epa.gov/sites/production/files/2015-08/documents/coverv14.pdf>.
- Valerio, Luis G. 2009. “In Silico Toxicology for the Pharmaceutical Sciences.” *Toxicology and Applied Pharmacology* 241 (3): 356–70. doi:10.1016/j.taap.2009.08.022.
- Vandenberg, Laura N. 2011. “Exposure to Bisphenol A in Canada: Invoking the Precautionary Principle.” *CMAJ: Canadian Medical Association Journal = Journal de l'Association Medicale Canadienne* 183 (11): 1265–70. doi:10.1503/cmaj.101408.
- Vandenberg, Laura N., Theo Colborn, Tyrone B. Hayes, Jerrold J. Heindel, David R. Jacobs, Duk-Hee Lee, Toshi Shioda, et al. 2012. “Hormones and Endocrine-Disrupting Chemicals: Low-Dose Effects and Nonmonotonic Dose Responses.” *Endocrine Reviews* 33 (3): 378–455. doi:10.1210/er.2011-1050.
- Vapnik, Vladimir N. 1998. *The Nature of Statistical Learning Theory*. New York, NY: Springer New York. doi:10.1007/978-1-4757-3264-1.
- Velasco-Santamaría, Yohana M., Bodil Korsgaard, Steffen S. Madsen, and Poul Bjerregaard. 2011. “Bezafibrate, a Lipid-Lowering Pharmaceutical, as a Potential Endocrine Disruptor in Male Zebrafish (Danio Rerio).” 105 (1–2). doi:10.1016/j.aquatox.2011.05.018.
- Victor-Costa, Anna Bolivar, Simone Miranda Carozzi Bandeira, André Gustavo Oliveira, Germán Arturo Bohórquez Mahecha, and Cleida Aparecida Oliveira. 2010. “Changes in Testicular Morphology and Steroidogenesis in Adult Rats Exposed to Atrazine.” *Reproductive Toxicology (Elmsford, N.Y.)* 29 (3): 323–31. doi:10.1016/j.reprotox.2009.12.006.
- Viel, Jean-François, Florence Rouget, Charline Warembourg, Christine Monfort, Gwendolina Limon, Sylvaine Cordier, and Cécile Chevrier. 2017. “Behavioural Disorders in 6-Year-Old Children and Pyrethroid Insecticide Exposure: The PELAGIE Mother-Child Cohort.” *Occupational and Environmental Medicine* 74 (4): 275–81. doi:10.1136/oemed-2016-104035.
- Vilella, Albert J, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. 2009. “EnsemblCompara GeneTrees: Complete, Duplication-Aware Phylogenetic Trees in Vertebrates.” *Genome Research* 19 (2): 327–35. doi:10.1101/gr.073585.107.
- Vink, S R, J Mikkers, T Bouwman, H Marquart, and E D Kroese. 2010. “Use of Read-across and Tiered Exposure Assessment in Risk Assessment under REACH--a Case Study on a Phase-in Substance.” *Regulatory Toxicology and Pharmacology : RTP* 58 (1): 64–71. doi:10.1016/j.yrtph.2010.04.004.

- Virtanen, H.E., and J. Toppari. 2007. "Epidemiology and Pathogenesis of Cryptorchidism." *Human Reproduction Update* 14 (1): 49–58. doi:10.1093/humupd/dmm027.
- Vonberg, David, Diana Hofmann, Jan Vanderborght, Anna Lelickens, Stephan Köppchen, Thomas Pütz, Peter Burauel, and Harry Vereecken. 2014. "Atrazine Soil Core Residue Analysis from an Agricultural Field 21 Years after Its Ban." *Journal of Environmental Quality* 43 (4): 1450–59. doi:10.2134/jeq2013.12.0497.
- Walker, William H. 2011. "Testosterone Signaling and the Regulation of Spermatogenesis." *Spermatogenesis* 1 (2). Taylor & Francis: 116–20. doi:10.4161/spmg.1.2.16956.
- Walschaerts, Marie, Audrey Muller, Jacques Auger, Louis Bujan, Jean-François Guérin, Dominique Le Lannou, André Clavert, Alfred Spira, Pierre Jouannet, and Patrick Thonneau. 2007. "Environmental, Occupational and Familial Risks for Testicular Cancer: A Hospital-Based Case-Control Study." *International Journal of Andrology* 30 (4): 222–29. doi:10.1111/j.1365-2605.2007.00805.x.
- Wang, Tao, and William E Rainey. 2012. "Human Adrenocortical Carcinoma Cell Lines." *Molecular and Cellular Endocrinology* 351 (1). NIH Public Access: 58–65. doi:10.1016/j.mce.2011.08.041.
- Waring, R H, and R M Harris. 2005. "Endocrine Disruptors: A Human Risk?" *Molecular and Cellular Endocrinology* 244 (1–2): 2–9. doi:10.1016/j.mce.2005.02.007.
- Weidner, I S, H Møller, T K Jensen, and N E Skakkebaek. 1998. "Cryptorchidism and Hypospadias in Sons of Gardeners and Farmers." *Environmental Health Perspectives* 106 (12): 793–96. <http://www.ncbi.nlm.nih.gov/pubmed/9831539>.
- Welter, Danielle, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, et al. 2014. "The NHGRI GWAS Catalog, a Curated Resource of SNP-Trait Associations." *Nucleic Acids Research* 42 (Database issue): D1001-6. doi:10.1093/nar/gkt1229.
- Westesson, Oscar, Mitchell Skinner, and Ian Holmes. 2013. "Visualizing next-Generation Sequencing Data with JBrowse." *Briefings in Bioinformatics* 14 (2): 172–77. doi:10.1093/bib/bbr078.
- "WHO | Global Assessment of the State-of-the-Science of Endocrine Disruptors." 2013. WHO. World Health Organization. http://www.who.int/gate2.inist.fr/ipcs/publications/new_issues/endocrine_disruptors/en/.
- Willemin, Marie-Emilie, Sophie Desmots, Rozenn Le Grand, François Lestremau, Florence A. Zeman, Eric Leclerc, Christian Moesch, and Céline Brochot. 2016. "PBPK Modeling of the Cis- and Trans-Permethrin Isomers and Their Major Urinary Metabolites in Rats." *Toxicology and Applied Pharmacology* 294 (March): 65–77. doi:10.1016/j.taap.2016.01.011.
- Willett, Peter. 2006. "Similarity-Based Virtual Screening Using 2D Fingerprints." *Drug Discovery Today* 11 (23–24): 1046–53. doi:10.1016/j.drudis.2006.10.005.
- Willett, Peter, John M. Barnard And, and Geoffrey M. Downs. 1998. "Chemical Similarity Searching." American Chemical Society. doi:10.1021/C19800211.
- Williams, Antony, and Valery Tkachenko. 2014. "The Royal Society of Chemistry and the Delivery of Chemistry Data Repositories for the Community." *Journal of Computer-Aided Molecular Design* 28 (10): 1023–30. doi:10.1007/s10822-014-9784-5.
- Winneke, Gerhard. 2011. "Developmental Aspects of Environmental Neurotoxicology: Lessons from Lead and Polychlorinated Biphenyls." *Journal of the Neurological Sciences* 308 (1–2): 9–15. doi:10.1016/j.jns.2011.05.020.

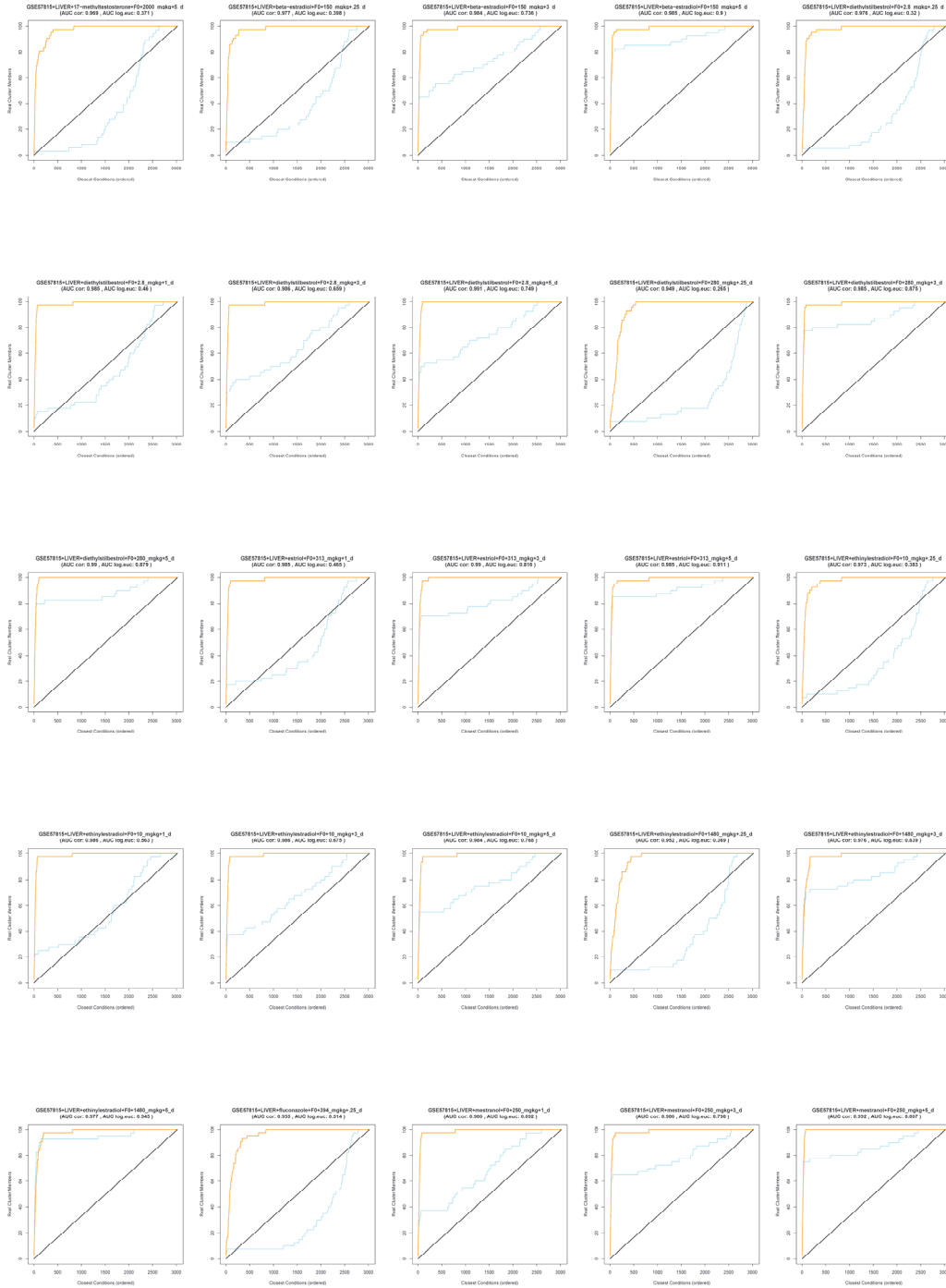
- Wirbisky, Sara E, Gregory J Weber, Maria S Sepúlveda, Changhe Xiao, Jason R Cannon, and Jennifer L Freeman. 2015. "Developmental Origins of Neurotransmitter and Transcriptome Alterations in Adult Female Zebrafish Exposed to Atrazine during Embryogenesis." *Toxicology* 333 (July): 156–67. doi:10.1016/j.tox.2015.04.016.
- Wurster, Charles F. 2015. *DDT Wars : Rescuing Our National Bird, Preventing Cancer, and Creating The Environmental Defense Fund*.
<https://global.oup.com/academic/product/ddt-wars-9780190219413?cc=fr&lang=en&>.
- Yang, Zheng Rong. 2004. "Biological Applications of Support Vector Machines." *Briefings in Bioinformatics* 5 (4): 328–38. <http://www.ncbi.nlm.nih.gov/pubmed/15606969>.
- Yates, Bethan, Bryony Braschi, Kristian A. Gray, Ruth L. Seal, Susan Tweedie, and Elspeth A. Bruford. 2017. "Genenames.org: The HGNC and VGNC Resources in 2017." *Nucleic Acids Research* 45 (D1): D619–25. doi:10.1093/nar/gkw1033.
- Yum, Taewoo, Sanghouck Lee, and Yunje Kim. 2013. "Association between Precocious Puberty and Some Endocrine Disruptors in Human Plasma." *Journal of Environmental Science and Health, Part A* 48 (8): 912–17. doi:10.1080/10934529.2013.762734.
- Zalko, Daniel, Carine Jacques, Hélène Duplan, Sandrine Bruel, and Elisabeth Perdu. 2011. "Viable Skin Efficiently Absorbs and Metabolizes Bisphenol A." *Chemosphere* 82 (3): 424–30. doi:10.1016/j.chemosphere.2010.09.058.
- Zhang, Boyang, Kunlun Huang, Liye Zhu, Yunbo Luo, and Wentao Xu. 2017. "Precision Toxicology Based on Single Cell Sequencing: An Evolving Trend in Toxicological Evaluations and Mechanism Exploration." *Archives of Toxicology* 91 (7): 2539–49. doi:10.1007/s00204-017-1971-4.
- Zhang, Wei, Hua Wang, Sonya W Song, and Gregory N Fuller. 2002. "Insulin-like Growth Factor Binding Protein 2: Gene Expression Microarrays and the Hypothesis-Generation Paradigm." *Brain Pathology (Zurich, Switzerland)* 12 (1): 87–94.
<http://www.ncbi.nlm.nih.gov/pubmed/11770904>.
- Zhao, Hao, Zhifu Sun, Jing Wang, Haojie Huang, Jean-Pierre Kocher, and Ligu Wang. 2014. "CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies." *Bioinformatics (Oxford, England)* 30 (7): 1006–7. doi:10.1093/bioinformatics/btt730.
- Zhou, Tong, Jeff Chou, Paul B Watkins, and William K Kaufmann. 2009. "Toxicogenomics: Transcription Profiling for Toxicology Assessment." *EXS* 99: 325–66.
<http://www.ncbi.nlm.nih.gov/pubmed/19157067>.
- Zhou, Z X, M V Lane, J A Kemppainen, F S French, and E M Wilson. 1995. "Specificity of Ligand-Dependent Androgen Receptor Stabilization: Receptor Domain Interactions Influence Ligand Dissociation and Receptor Stability." *Molecular Endocrinology (Baltimore, Md.)* 9 (2): 208–18. doi:10.1210/mend.9.2.7776971.
- Zlatnik, Marya G. 2016. "Endocrine-Disrupting Chemicals and Reproductive Health." *Journal of Midwifery & Women's Health* 61 (4): 442–55. doi:10.1111/jmwh.12500.
- Zoeller, R. T., and J. Rovet. 2004. "Timing of Thyroid Hormone Action in the Developing Brain: Clinical Observations and Experimental Findings." *Journal of Neuroendocrinology* 16 (10). Blackwell Science Ltd: 809–18. doi:10.1111/j.1365-2826.2004.01243.x.
- Zoeller, R. Thomas, T. R. Brown, L. L. Doan, A. C. Gore, N. E. Skakkebaek, A. M. Soto, T. J. Woodruff, and F. S. Vom Saal. 2012. "Endocrine-Disrupting Chemicals and Public Health Protection: A Statement of Principles from The Endocrine Society." *Endocrinology* 153 (9): 4097–4110. doi:10.1210/en.2012-1422.

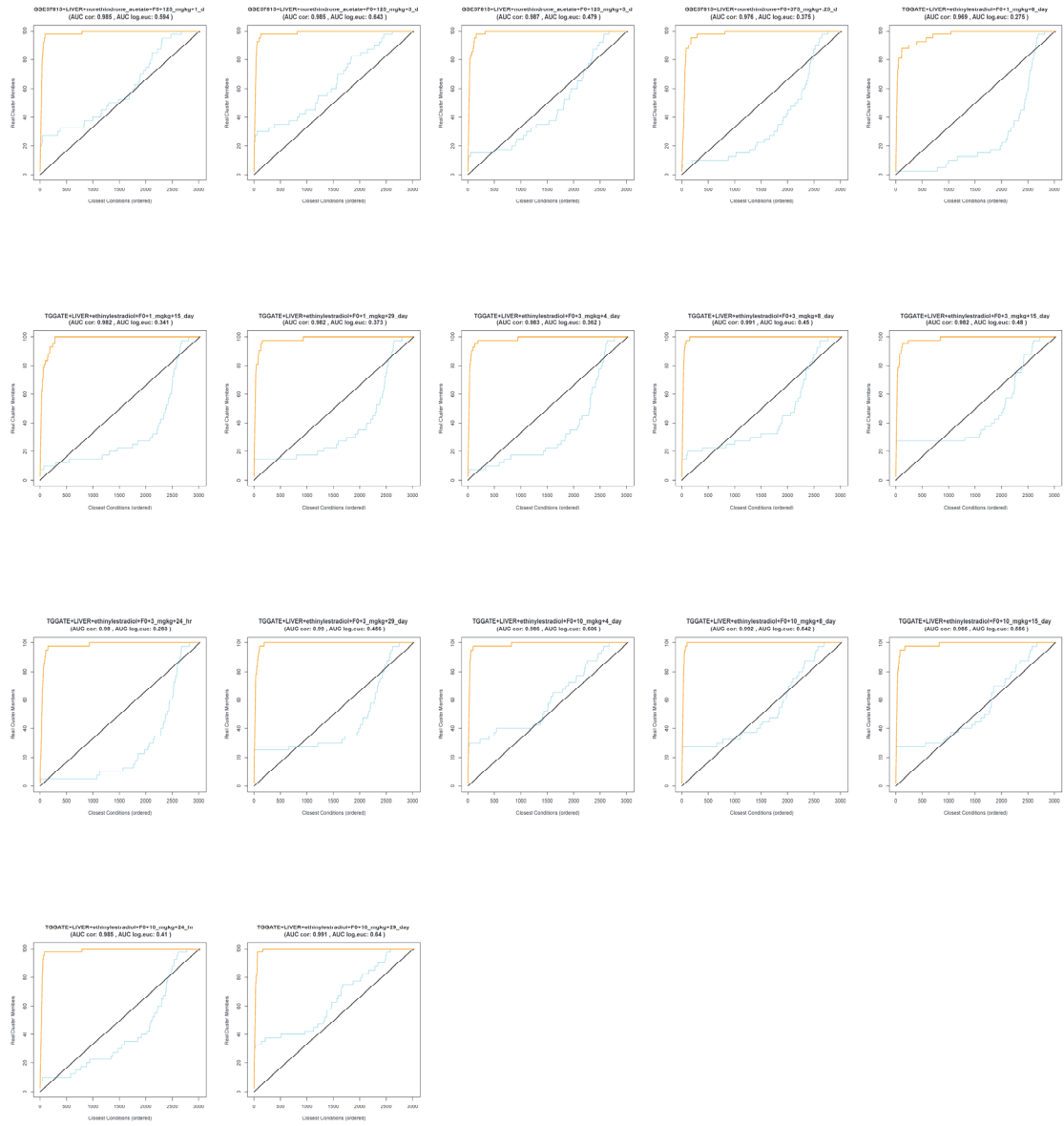
Zuoshuang, Xiang,, Mungall; Chris, Rutterberg; Alan, and He Yongqun. 2011. “Ontobee: A Linked Data Server and Browser for Ontology Terms.” *ICBO: International Conference on Biomedical Ontology* 1: 1–3.

Annexes

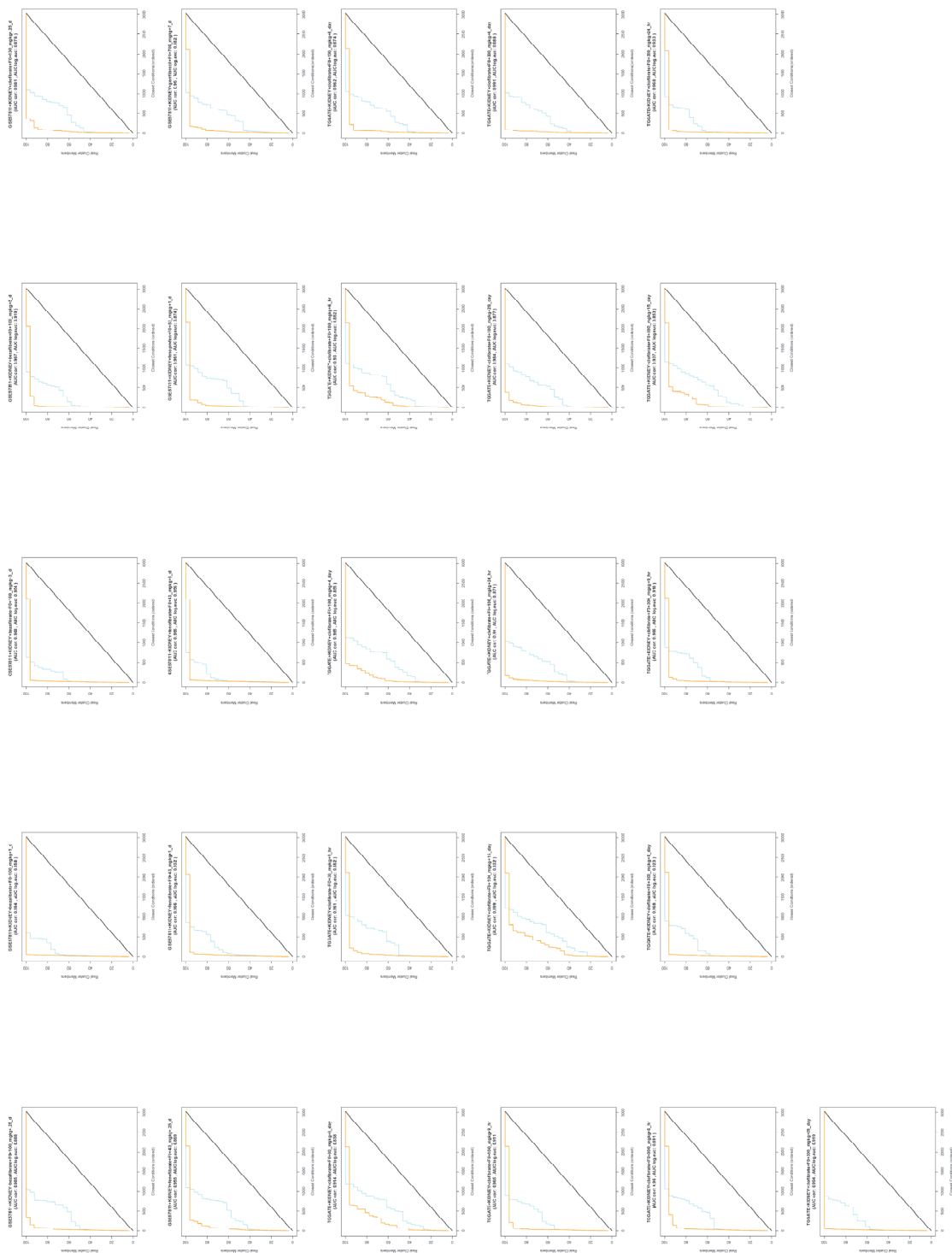
1. Conditions expérimentales des groupes sélectionnés

1.1. Conditions expérimentales du groupe 1

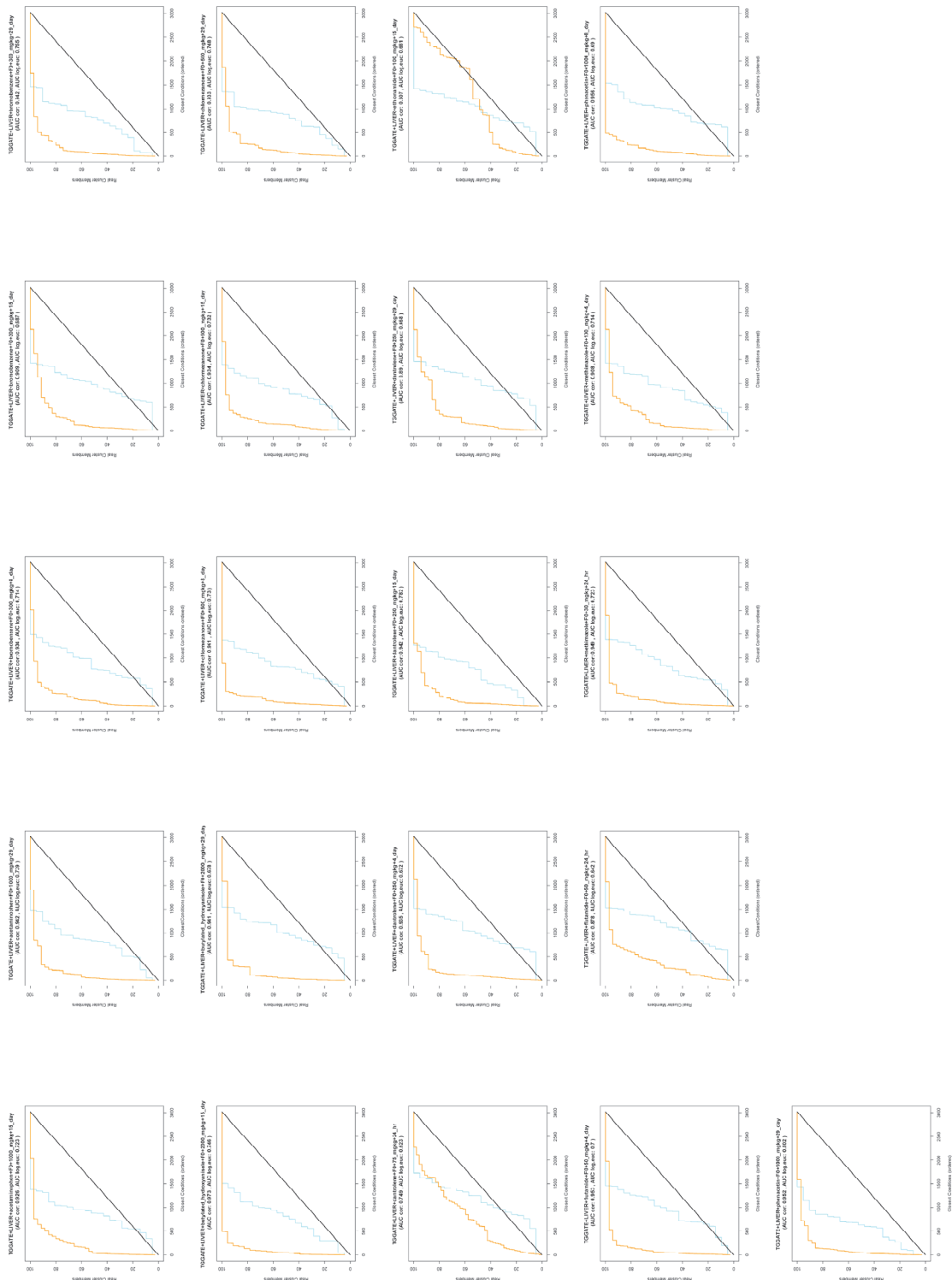




1.2. Conditions expérimentales du groupe 2



1.3. Conditions expérimentales du groupe 3



VU :

VU :

**Le Directeur de Thèse
Doctorale**

(Nom et Prénom)

Le Responsable de l'École

VU pour autorisation de soutenance

Rennes, le

Le Président de l'Université de Rennes 1

David ALIS

VU après soutenance pour autorisation de publication :

Le Président de Jury,
(Nom et Prénom)

Résumé

L'un des plus importants défis de la toxicologie est de pouvoir extrapoler les résultats issus des différentes phases de l'analyse du risque sanitaire à partir de systèmes expérimentaux vers les populations humaines. Dans ce contexte, les techniques globales dites "omiques" sont de plus en plus utilisées pour caractériser les différents états des systèmes biologiques. Ainsi, la toxicogénomique permet non seulement d'étudier les mécanismes d'action des composés, d'identifier des marqueurs d'exposition, mais aussi de générer des signatures moléculaires à potentiel prédictif. En effet, des composés ayant des signatures moléculaires semblables ont également de forts risques de présenter les mêmes effets toxicologiques.

L'objectif de cette thèse est d'appliquer ce concept de manière systématique, en explorant les données publiées et disponibles dans les banques dédiées à la toxicogénomique via des modèles statistiques multivariés. Ces analyses ont pour objectif de permettre le regroupement et donc la classification des composés sur la base des gènes dont ils affectent l'expression. L'appartenance de produits toxiques bien caractérisés aux classes ainsi constituées permet alors d'émettre des hypothèses quant à la toxicité des autres composés. Un jeu de données quantitatives intégrant 18 études réalisées avec la même technologie de puce à ADN et chez une seule espèce a été assemblé. De ce jeu de données, 3022 signatures toxicogénomiques correspondant à 452 composés différents ont été obtenues. Des approches de classification non supervisées afin de définir des classes de traitements induisant des altérations transcriptionnelles proches ont été mises en place. 95 et 104 classes ont été obtenues en fonction des méthodes utilisées. Finalement, une attention toute particulière a été portée sur les potentiels nouveaux perturbateurs endocriniens et xénobiotiques reprotoxiques sur-représentés dans trois classes spécifiquement. 22 composés sont en cours de test sur une lignée cellulaire humaine exprimant les enzymes de la stéroïdogénèse (NCI-H295R) pour évaluer leur potentiel effet perturbateur endocrinien.

L'ensemble de ce travail a ainsi permis de démontrer la pertinence de nos approches de toxicogénomique pour la prédiction des effets toxiques de composés testés dans d'autres organes et/ou chez d'autres espèces. Il se poursuit à l'heure actuelle par la mise en place d'une base de données, *TOXsIgN*, permettant l'hébergement et l'accès à l'ensemble de signatures de toxicogénomique générées dans ce projet. De même, mon travail a également permis la mise en place de plusieurs outils dédiés à la toxicologie prédictive et à la visualisation de données, tels que le navigateur de génomes comme le *ReproGenomics Viewer* (RGV).