

# UNIVERSITÉ DE TOURS

## ÉCOLE DOCTORALE : *SSBCV*

**ÉQUIPE : BINGO - Biologie Intégrative des Gonades**

UMR0085 Physiologie de la Reproduction et des Comportements INRAe Centre Val de Loire

**THÈSE** présentée par :

Floriane PICOLO

soutenue le : **08 décembre 2023**

pour obtenir le grade de : **Docteure de l'Université de Tours**

Discipline/Spécialité : Bioinformatique

**Étude de l'évolution des gènes codant les  
protéines de transduction des signaux  
intracellulaires chez les animaux**

**THÈSE dirigée par :**

M MONGET Philippe  
M PIÉGU Benoît

Directeur de recherche, Inrae  
Ingénieur de recherche, Inrae

**RAPPORTEURS :**

M BOBE Julien  
M FARAUT Thomas

Directeur de recherche, Inrae  
Chargé de recherche, Inrae

**JURY :**

Mme DUITTOZ Anne (*Présidente du jury*)  
Mme LANDES Claudine  
M BOBE Julien  
M FARAUT Thomas  
M MONGET Philippe  
M PIÉGU Benoît

Professeure, Université de Tours  
Professeure, Université d'Angers  
Directeur de recherche, Inrae  
Chargé de recherche, Inrae  
Directeur de recherche, Inrae  
Ingénieur de recherche, Inrae

# Remerciements

Cette thèse a été réalisée sous la direction de Philippe Monget au sein de l'équipe Biologie INtégrative des GOnades (BINGO) et sous l'encadrement de Benoît Piégu de l'unité Physiologie de la Reproduction et des Comportements (PRC) à l'INRAE de Nouzilly et grâce aux financements de l'Université de Tours.

Je remercie donc l'ensemble des membres de la direction d'unité et tout particulièrement Sophie Mary, Anne Mychak, Ghyslaine Ploux et Sandra Cavalie pour leur aide administrative.

Je tiens principalement à remercier les membres du jury d'avoir accepté d'évaluer mon travail, Julien Bobe et Thomas Faraud en tant que rapporteurs ainsi que Anne Duittoz et Claudine Landes en tant qu'examinatrices.

Un profond merci à mon directeur de thèse Philippe Monget pour avoir cru en moi dès le début de mon premier stage en 2016 puis en 2019 et de m'avoir donné cette chance de réaliser une thèse sur un sujet passionnant. Tu as été un réel soutien sur le plan scientifique bien évidemment, pour tous les conseils durant ces trois années, mais surtout pour ton efficacité sans nom dans les retours que ce soit sur les articles ou ce manuscrit. Et tu as également été une bonne épaule pour moi, la doctorante un peu (*beaucoup*) fragile émotionnellement, et qui n'a pas toujours été facile à gérer. Et je tiens également à te dire que je suis ravie que tu ne viennes pas à ma soutenance avec une perruque bleu-blanc-rouge malgré une défaite douloureuse de la France au Rugby.

Je souhaite également remercier mon encadrant Benoît Piégu, pour le soutien et les déblocages sur R. Merci d'avoir insisté si **lourdement** pour que je sauvegarde mes avancements sur GIT, tu auras finalement eu raison de moi, malgré ma ténacité à te tenir tête. On se souviendra des petites blagues aux JOBIM 2023, mais aussi de toutes celles durant ces trois ans. Et un merci tout particulier pour ces parties de ping-pong mailistique durant la rédaction.

Merci à Elis, le chef de notre équipe BINGO. Je n'aurais finalement jamais goûté aux chocolats que tu prétends donner à tous les membres de l'équipe, venant pour tes superbes conseils dignes d'un psychologue. Merci d'être toi, et de m'avoir intégré dans l'équipe.

Merci à tous les membres de l'équipe BINGO : Sébastien Elis, Philippe Monget, Marie-Émilie Lebachelier, Sophie Fouchécourt, Virginie Maillard, Peggy Jarrier-Gaillard, Pascal

Papillier, Rozenn Dalbies-Tran, Véronique Cadoret, Sandrine Fréret, Ghylène Goudet, Maria-Teresa Pellicer-Rubio, Catherine Taragnat, Svetlana Uzbekova, Laetitia Corset, Cécile Douet, Stéphanie Martinet, Charline Pontlevoy, Claire Vignault, Ophélie Téteau, Alice Desmarchais et Anna Grandchamp.

Un merci tout particulier à Sophie Fouchécourt qui a été la première à croire en moi et à me donner cette chance lors d'un stage en DUT en 2016 alors que je n'étais qu'une novice en bio-informatique. Merci de m'avoir accueilli et de m'avoir fait confiance.

Un grand merci également à Michel Laurin et Jérémie Bardin du MNHN de Paris avec qui nous avons grandement collaboré avec un immense plaisir. Bien que l'on ne se soit jamais rencontré, j'ai adoré travailler avec vous.

Un merci à l'équipe de PIXANIM, Valoch, Ana Paula ou AP, Dany, Mimi, Cholet et Jérôme de m'avoir accueillie dans leur bureau pour prendre le café, se raconter nos péri-péties, rigoler, faire des compétitions de pédantix et peindre même à l'occasion ! Chaque instant était un plaisir avec vous.

Je remercie aussi les stagiaires de l'équipe Maoui alias le meilleur stagiaire dont je me souviendrais toujours de l'histoire du vélo dans le bus, et tes skills éclatés au sol sur Excel. Et Mapâte la meilleure piou-piou devenue PIOUUU maintenant avec qui on a beaucoup rigolé de ta drôle de passion pour les diverticules, et d'ailleurs, je reste fan de tes slides animés.

Je vais remercier également les zozos du fond du couloir qui font énormément de bruits sans fermer leur porte, les doctorants de l'équipe SENSOR bien évidemment, dont le Bourdz mon partenaire de crêpe et de chouchen, Loïse celle qui a copié ma date de soutenance, je ne serais pas là pour te voir, mais force ! Et entre autres, Mathy que j'ai allégrement embêté en cette fin de thèse.

À Marie et Coline, mes co-thésardes préférées, qui m'ont sorti de l'isolement après 1 an de thèse cloitrée à la maison puis dans un bureau isolé au sous-sol. Marie, tu as été d'un soutien immense durant cette thèse. Merci pour tous tes conseils, mais aussi pour toutes ces conversations dans le bureau, et tous ces BeReal ensemble. Oui, c'est Michel, tu donnes pas de nouvelles. Et Coline la meilleure thésarde du monde, un modèle d'exemplarité ! Sous tes petits mots fleuris, j'ai su déceler tout l'amour que tu m'apportais, et finalement... bon, je t'aime aussi un peu en retour. Avec toi, c'est l'amour vache (je suis hyper drôle) mais j'ai passé que des bons moments à gueuler avec toi (*et sur toi*). D'ailleurs, je t'offre ma voiture, parce que finalement tu l'as plus conduite que moi, et on sait pourquoi. Finalement, à linera, on ne se fait pas que des collègues, on se fait aussi des bêtes de copines.

Un énorme merci aussi à Mimi, la plus fun de ma vie, qui prend si soin de moi. Tu m'auras mise une musique par jour dans la tête, tu m'auras épuisée de ton trop pleins d'énergie et de tous tes uppercuts, mais tu vas me manquer quand tu seras de l'autre côté de l'Atlantique. J'ai adoré te faire des chasses aux trésors et dessiner sur ton bureau pour

voir mon bureau multicolore en retour. C'était de bonne guerre. Et un big up à Pouf, Joël et sa poire.

Et puis 3 ans, c'est aussi partager son quotidien avec ses voisins ! Merci aux Velpotes : Romain, Pierre, Valentin, Léa et Mathilde. Vous avez rythmé ma vie en partageant des soirées jeux, des pétanques ou des dîners, et vous m'avez même remotivé à courir ! On a vécu des moments improbables ensemble comme l'étrangleur ou Bertrand Desplantes qui ont rempli ma boite à souvenirs. Et un petit mot pour Valentin, mon copain de DUT, merci d'avoir repris contact, merci à nos soirées télés, et de m'avoir fait rencontrer Doriane, cette merveilleuse personne que je vois maintenant plus que toi finalement... Vous êtes au top les velpotes.

J'ai également rencontré de formidables personnes grâce à l'Association des Doctorants de Tours dont j'ai été membre pendant mes 3 années et grâce à qui j'ai rencontré mes copaines de Toursnez Ménage qui m'ont chaleureusement accueilli dans leurs bureaux : Antonin, Yopo, Léa, Thomas, Romain, Igor, Yegor, Guillaume, Théo, Adam, Pierre, Sylvain, Merve, Maxime et Marion ! Que des numéros un dans cette team ! Ça aura été un plaisir de vous gronder chaque jeudi pour que vous veniez faire la fête !

Dans cette même veine, j'aimerais faire un aparté pour Romain, mon voisin et maintenant meilleur ami, j'ai passé de merveilleux moments avec toi, à rigoler et faire les andouilles (*qu'on adore*). Puis notre trio de l'enfer avec Thomas, à chanter sur les quais de Loire ou juste à se raconter nos vies sur la place Velpeau jusqu'à pas d'heure. Et aussi merci de m'avoir fait rencontrer Sam, notre artiste qui m'a aidé à décompresser pendant la rédaction avec nos jeux de coop. Merci !

Et puis je suis obligé de remercier dignement mes parents adoptifs du Shamrock, Annie et Fafa ! Deux formidables personnes, à qui je souhaite un bon départ vers de nouveaux horizons. Annie, merci d'avoir été ma maman à Tours, toujours à nous vouloir nous faire plaisir et à prendre soin de nous. Et que vive longtemps le PICOLO Grande !

J'aimerais également remercier du fond de mon cœur, le serveur discord des doctorants « PhD Student » que j'ai rejoint le premier jour de thèse, et que je n'ai jamais lâché depuis. De nombreuses amitiés se sont créées grâce à cette plateforme en ligne, et on a pu se soutenir et rire pendant la pandémie, mais bien plus après notamment à travers les IRL, mais aussi en comptant l'heure (*ça n'a pas de sens*) mais merci aux horlogistes.

Un grand merci à Anana, ma plus belle rencontre sur ce serveur, qui au premier abord n'était qu'une camarade de thèse, mais qui s'est transformé en amitié si intense. Je ne compte plus les allers-retours jusqu'à Lille pour passer un week-end avec toi et les dinos LoloR et Ouistiti. Merci d'être la personne que tu es, tu m'as fait grandir sur bien des sujets et notamment la communication. Finalement, si j'arrive à m'ouvrir dans ses remerciements, c'est peut-être grâce à toi.

Un merci à Galaad aussi, j'ai adoré nos échanges, nos tierlists, et nos jeux (*achetés pour ne plus y toucher*), notre projet de tout plaquer pour lancer une chaîne Twitch n'aura

finalement pas eu lieu, mais on avait un si beau concept. Je le garde en tête. Et un gros signe JuL à notre groupe Salut l'Ariège, et à nos AG exceptionnelles pour organiser nos vacances.

Un merci plus général aux équipes de modération du serveur « PhD Student ». En deux ans de modération, j'en ai vu des camarades, et j'ai adoré travailler avec chacun d'entre vous. Un jour, je partirais, mais je suis sûre que mes poulains sauront rédiger des comptes rendus qui me rendront digne (*j'en fais des tonnes*). Courage !

Et puis, jamais très loin, j'ai les copains du master, on ne s'est pas vu normalement, mais c'était pour profiter encore plus lorsque l'on se retrouvait aux BIGDAY, aux JOBIM ou lors des soutenances de thèses. À ma meilleure amie d'étude supérieur, Manon, qui est maintenant Docteure. Tu es la meilleure ne l'oublie pas.

Je remercierai également mes amies du lycée et principalement Angèle et Laurie, mes deux meilleures amies depuis si longtemps et avec qui on ne s'est jamais lâché. Chaque instant avec vous vaut le dé*Tours* (jeu de mot, parce que je reste l'humour du groupe). J'espère qu'un jour les gars (*et nous*) finiront par grandir et on pourra partir sereinement au ski. Enfin, quoique, ça fait tellement de bons souvenirs.

Et enfin ma famille, je ne sais si je trouverai les mots pour vous remercier. À mon père et ma belle-mère qui m'ont toujours écouté me plaindre sans forcément comprendre. À ma maman parfaite et Sergio qui pensent encore que je vais bientôt devenir Avocate, désolée c'est toujours pas ça. À mon frère à qui je souhaite de réussir, tu as des mains en or ma breur. À mon arrière-grand-mère décédée, et à mon arrière-grand-père qui nous a également quitté durant ma rédaction. À mon grand-père et à ma merveilleuse grand-mère. Ça n'a jamais été facile de se dire je t'aime, mais je crois que je peux au moins l'écrire. **Je vous aime.**

Et puis un petit merci, à mon chat, Nuggets. Ce gros sac à prout a rythmé mes nuits et mes insomnies. Un enfant capricieux, mais que l'on aime de tout son cœur.

Alors voilà que s'achève 3 années de ma vie. Quand je repense à Philippe le 15 novembre 2020 me disant « tu verras, 36 mois ça passe très vite » et moi lui riant au nez, finalement, tu avais raison. En 36 mois, j'ai eu trois covid compliqués, dont un durant ma soutenance, une grippe, un accident de voiture, et je suis aussi passée sous une voiture. J'ai eu le feu dans mon appartement et deux dégâts des eaux. Je me suis fait voler mes affaires sur les bords de Loire et j'ai eu un étrangleur dans mon immeuble. Mais j'ai également j'ai fait partie de deux associations incroyables, créé un jeu de piste sur la ville de Lyon et créé un cocktail au Shamrock de Tours, organisé deux IRL géantes pour les doctorant.es en France, couru 10 km et battu mon propre record.

***Mais surtout, j'ai réussi à écrire cette thèse.***

Merci à vous.

Bravo à moi.



---

## Table des matières



---

## Table des figures

# Introduction

Si l'évolution peut avoir plusieurs définitions dans le dictionnaire, elle a une définition très spécifique dans l'univers de la biologie. En effet, l'évolution se réfère à la modification progressive du monde vivant au fil du temps. L'évolution est un mécanisme complexe et permet la diversité de la vie sur Terre. Elle peut se caractériser par plusieurs sous-notions comme la notion d'hérédité, d'adaptation, de sélection, de coévolution, de mutation ou encore de spéciation. Ces grandes notions vont être importantes pour la suite de l'introduction : la génétique, l'étude des gènes, et génomique, celle des génomes. Dans les prochains chapitres introductifs, nous détaillerons ce que sont les gènes et génomes, la diversité animale, et brièvement la communication cellulaire.

## 1.1 Gènes et génomes

Les gènes sont des fragments d'ADN (acide désoxyribonucléique) qui comportent l'information génétique de chaque organisme vivant. L'ADN mitochondrial est stocké principalement au niveau du noyau de la cellule pour les organismes eucaryotes (à la différence des organismes procaryotes, donc sans noyau où l'ADN est retrouvé concentré en région que l'on appelle le nucléoïde). Les gènes sont transmis d'une génération à une autre, mais sont soumis à des modifications aussi appelées « mutations » de leur séquence nucléotidique (succession d'éléments : adénine (A), cytosine (C), guanine (G) et thymine (T)). Pour être traduit en protéine, la séquence nucléotidique d'un gène se lit par codon, une suite de 3 nucléotides. Chaque combinaison de codon correspond à un acide aminé. Il existe 20 acides aminés communs à l'ensemble du vivant (**ambrogelly\_natural\_2007**). Une suite d'acides aminés commençant par un codon initiateur et se finissant par un codon terminal deviendra sera dans un premier temps transcrit en ARNm (acide ribonucléique messager) puis traduit en protéine par le biais de la traduction réalisée par un ribosome.

La transcription a lieu dans le noyau de la cellule, et transcrit donc la séquence nucléotidique en séquence complémentaire. Il est important de préciser qu'un même gène peut conduire à une multitude de protéines (**breathnach\_organization\_1981**). En effet, les gènes sont constitués de sous-parties : d'un promoteur (partie initiatrice de

la transcription), d'exons (parties codantes de l'ADN, elles vont aider à déterminer la structure de la protéine) et d'introns (parties non codantes de l'ADN, elles vont jouer un rôle dans la régulation de l'expression génique) ([scherrer\\_pre-messenger\\_1979](#); [keren\\_alternative\\_2010](#); [shaul\\_how\\_2017](#)). Chez les eucaryotes, avant que l'ARNm ne soit traduit en protéines, il va être édité et ce phénomène s'appelle la maturation. Les introns vont être épissés, et les exons conservés. Un même gène peut avoir différentes combinaisons d'exons assemblés, ce qui peut conduire à avoir plusieurs protéines différentes (Figure ??).

Comme on a pu le voir, les gènes sont des petits éléments complexes, mais indispensables à un organisme. L'ensemble des gènes constitue le génome.

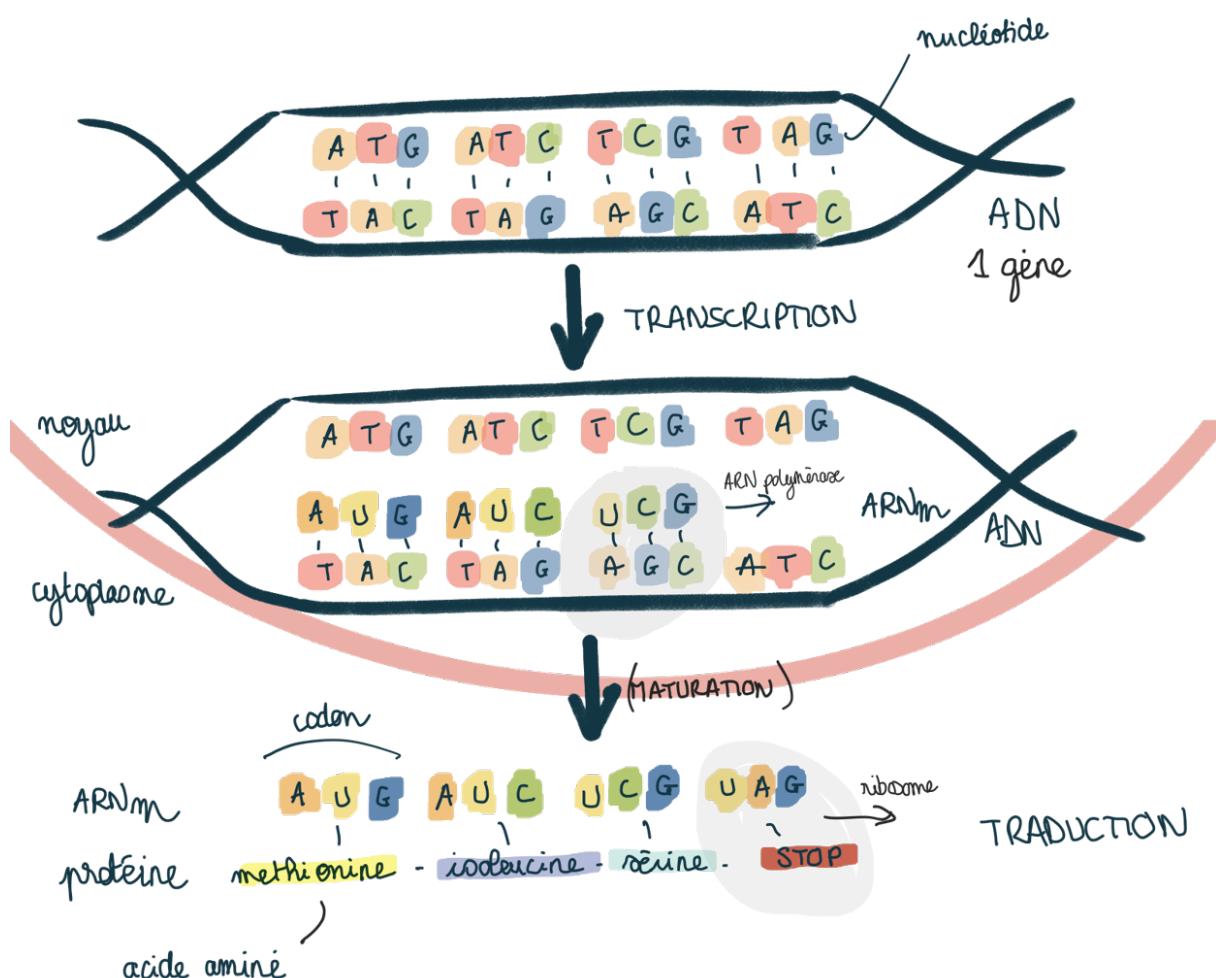


FIGURE 1.1 – Schéma d'un gène à une protéine

Légende : Les différentes étapes d'un gène à une protéine sont représentées à l'exception de la maturation (épissage alternatif).

### 1.1.1 Évolution des gènes : mutations, délétions, insertions

L'ADN peut subir des changements de différentes manières. Tout d'abord, il y a les mutations, qui peuvent être spontanées, se produisant lors de la transmission de l'information génétique à la descendance, ou environnementales, résultant d'altérations dues à des facteurs extérieurs tels que les virus et le soleil, entre autres. Ces mutations affectent la séquence de nucléotides et peuvent entraîner des conséquences plus ou moins graves en fonction de leur nature.

Une mutation peut être une substitution d'un nucléotide par un autre. En fonction du nucléotide touché, elle peut avoir plus ou moins de conséquences. Le code génétique est doté de quelques redondances concernant la traduction des acides aminés. Par exemple, la sérine est un acide aminé codé par 6 codons différents (UCU, UCC, UCA, UCG, AGU et AGC). Si la séquence est initialement UCG, et que la substitution a lieu sur le 3ème nucléotide de la séquence : il n'y aura pas d'impact sur la traduction du codon, quelle que soit la substitution. Cette mutation est dite silencieuse. Si la substitution a lieu sur le 2ème nucléotide, de telle façon à former UAG, on obtient donc un codon stop, ce qui veut dire que le ribosome n'ira pas au-delà lors de la traduction. La protéine ne sera donc pas celle initialement prévue, possiblement tronquée et non fonctionnelle. Cette mutation est dite non-sens. Et enfin si la substitution a lieu sur le 1ème nucléotide, de telle façon à former GCC, on obtient un nouvel acide aminé qui est lalanine. Cette mutation est dite faux sens car la protéine finale sera impactée par ce changement d'acide aminé (Figure ??).

Une mutation peut aussi prendre la forme d'une insertion ou d'une délétion, c'est-à-dire l'ajout ou la suppression d'un nucléotide dans la séquence d'ADN. Étant donné que la séquence se lit en codons, cela perturbera le "cadre de lecture" si un ou plusieurs nucléotides (non multiple de 3) sont ajoutés ou supprimés. Le cadre de lecture détermine comment les codons seront lus par le ribosome lors de la traduction (Figure ??).

Il existe également la recombinaison génique qui est un déplacement d'une région d'ADN vers une autre. Lors de ces recombinaisons, il se peut que la séquence soit également dupliquée ([sasaki\\_genome\\_2010](#); [stewart\\_homologous\\_2022](#); [syeda\\_recombination\\_2014](#)). Des copies de gènes ou des copies de séquence sont alors ajoutées dans la séquence, que l'on appelle des duplications segmentales. Lorsqu'il s'agit de gène entièrement dupliqué, différents scénarios peuvent avoir lieu. Le gène dupliqué peut garder la fonction initiale du gène original. Le gène peut acquérir une nouvelle fonction proche, voire une spécialisation (spatio-temporel) au gène original. Mais il peut également se pseudogéniser, c'est-à-dire devenir non fonctionnel par une altération de sa séquence.

Les mutations ont donc un effet direct sur la séquence nucléotidique, mais également sur la protéine et la fonctionnalité de celle-ci, effet qui peut être avantageux ou désavantageux ([long\\_origin\\_2003](#)). Les duplications sont également un moteur de l'évolution.

L'exemple de la famille des globines peut en attester (**weatherall \_ molecular \_ 1976**). En effet, les duplications des globines ont permis aux vertébrés de s'adapter à différents environnements ou à des contraintes physiologiques.

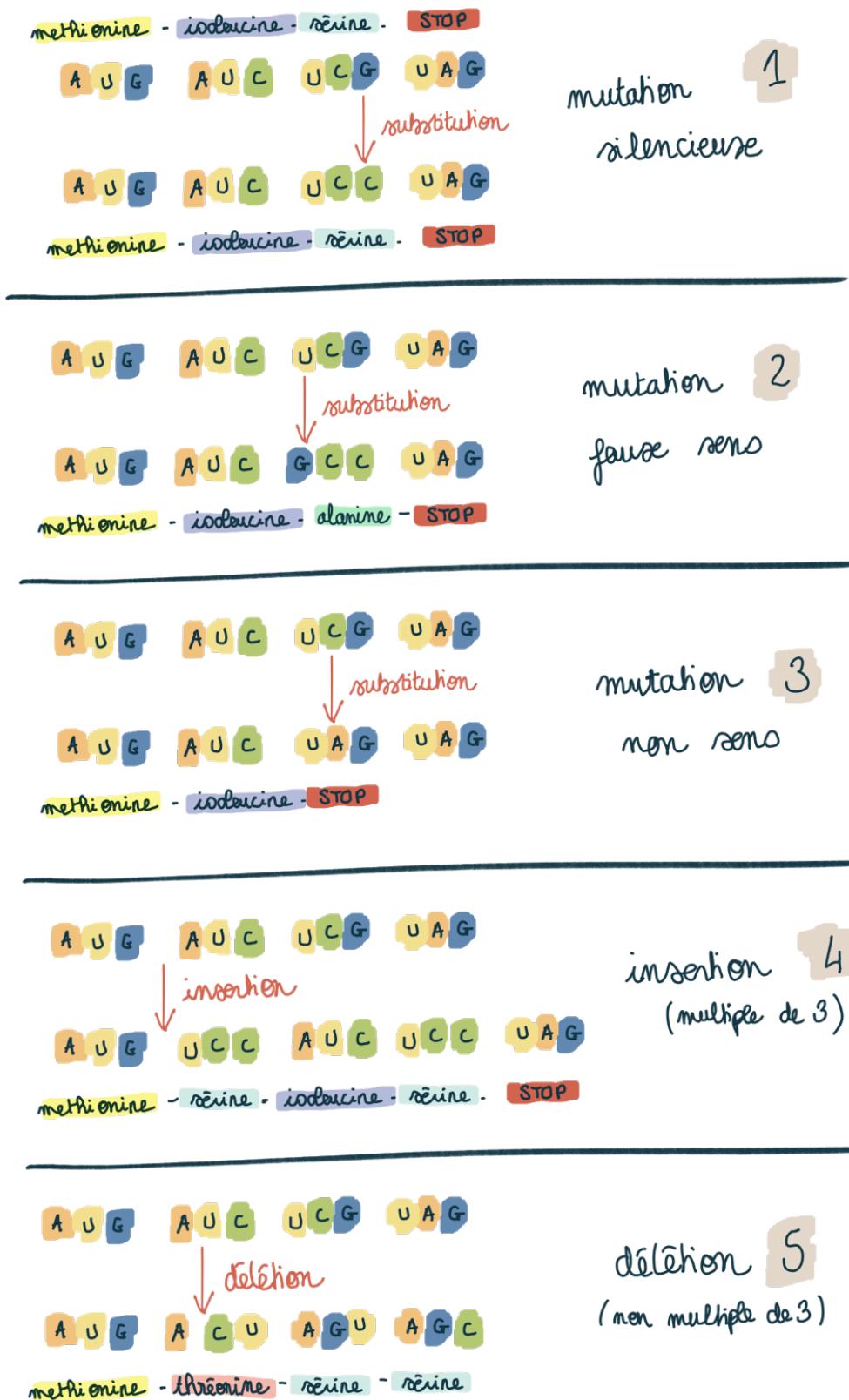


FIGURE 1.2 – Différents types de mutations

Légende : Représentation des mutations silencieuses, faux sens, non-sens, des insertions et délétion à partir d'une même séquence initiale.

Les duplications segmentales sont des événements à échelle restreinte, mais il existe également une autre forme de duplication, la duplication complète du génome.

### 1.1.2 Duplication complète de génome

La duplication de génome (Whole Génome Duplication en anglais, WGD) est un événement de polyploidisation, c'est-à-dire une duplication du nombre de chromosomes. La polyploidisation est très courante dans le règne des plantes, puisque environ 70% des angiospermes ont eu au moins un événement de polyploidisation (**masterson\_stomatal\_1994**; **soltis\_polyplody\_2009**).

Les premières études démontrant des phénomènes de duplication ancestrale de génome chez les plantes date des années 2000 chez *Arabidopsis Thaliana* (**blanc\_recent\_2003**; **bowers\_unravelling\_2003**; **the\_arabidopsis\_genome\_initiative\_analysis\_2000**) et concernant le règne des animaux, les duplications de génome sont mises en évidence chez les vertébrés à partir de 1994 (**holland\_gene\_1994**; **nakatani\_reconstruction\_2007**). Le groupe des vertébrés est notamment marqué par la succession de deux duplications complètes de génome par rapport aux invertébrés (**dehal\_two\_2005**). Ce double événement est d'ailleurs probablement impliqué dans la divergence des espèces de vertébrés. Le groupe des vertébrés ne comptant « que » 72 000 espèces (contre environ 1 000 000 d'invertébrés d'après la Classification phylogénétique du Vivant (**lecointre\_classification\_2016**)) regroupe, pour autant, une multitude d'espèces de tailles et de morphologies, d'environnements ou de modes de reproduction complètement différentes.

D'autres groupes ont également vécu des duplications de génome supplémentaires, comme le groupe des poissons téléostéen (**braasch\_polyplody\_2012**; **meyer\_2r\_2005**), dont celui des salmonidés et des carpes (**lien\_atlantic\_2016**; **xu\_allotetraploid\_2019**) que nous verrons par la suite plus en détail.

Une famille de gènes a notamment permis de mettre en lumière les duplications de génomes : la famille des Hox (venant de « *homeobox* »). Les gènes Hox sont des facteurs de transcription présents chez la quasi-totalité des bilatériens (*Bilateria*). Leur organisation génomique est relativement complexe, c'est-à-dire localisés par bloc de plusieurs gènes à la suite. Par exemple, la drosophile compte 8 gènes en 2 complexes, et l'homme 39 gènes organisés en 4 complexes, chacun est subdivisé en 13 paralogues (gènes Hox1 jusqu'à Hox13) (**hoeegg\_hox\_2005**; **meyer\_vertebrate\_1999**; **rux\_hox\_2017**). Chez les téléostéens, on compte pas moins de 63 gènes Hox (**meyer\_vertebrate\_1999**; **stellwag\_hox\_1999**) (Figure ??). De ces premières découvertes, il a été question de trouver l'origine de ces multiples duplications, et de déterminer s'il s'agissait de dupli-

tions segmentales ou de duplications de génome.

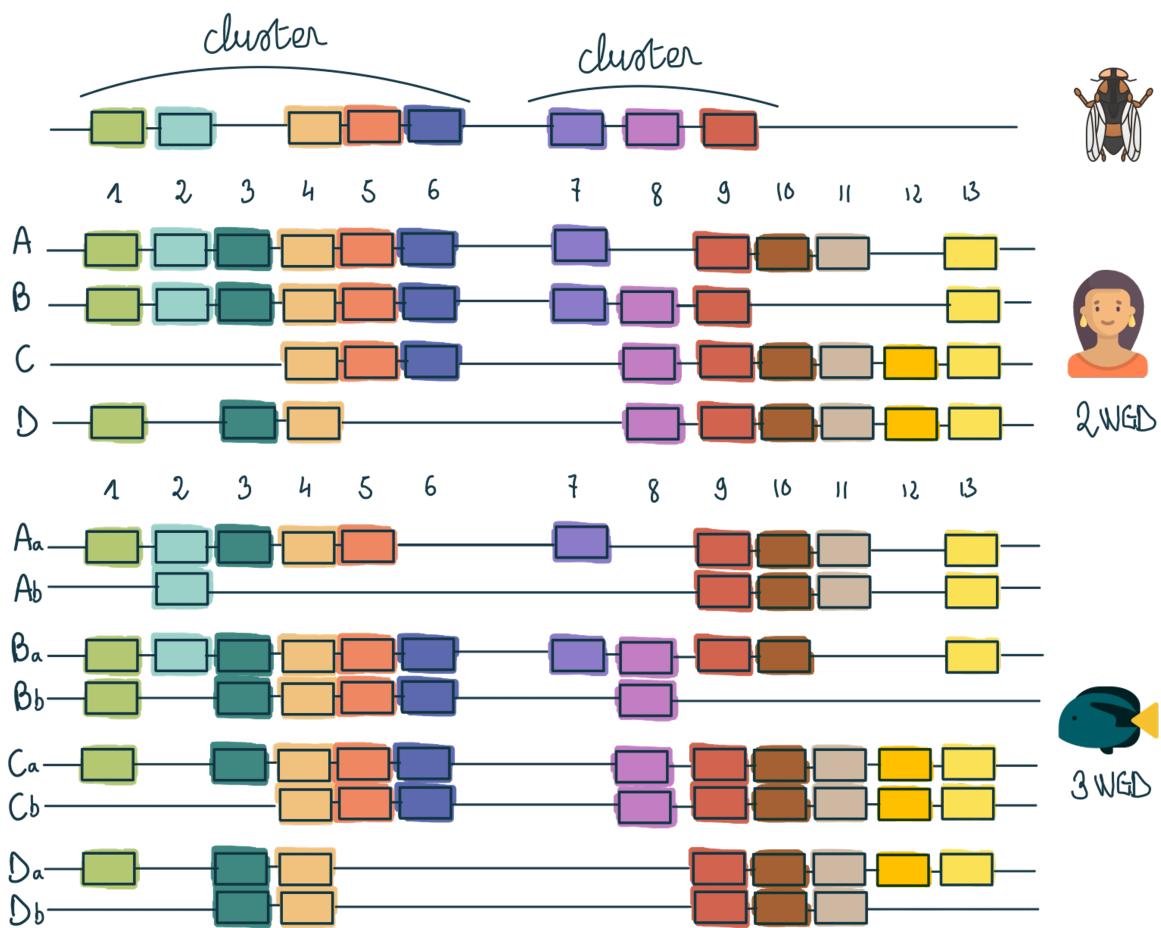


FIGURE 1.3 – Organisation des gènes Hox chez la drosophile, les mammifères et les téléostéens

Légende : Chaque carré représente un gène Hox. Pour la drosophile, les clusters vont de 1 à 6 puis de 7 à 13. Pour les mammifères et les téléostéens, ce seront préférentiellement les différentes lettres A à D. Les gènes partageant la même couleur sont des orthologues. Entre la drosophile et les mammifères, le gène 2 s'est subdivisé en 2 et 3, et le gène 9 s'est subdivisé en 9 à 13. Les mammifères qui sont un groupe d'espèces à 2 WGD (duplication de génome complet) comptent environ 39 gènes, et les téléostéens (3 WGD) comptent environ 63 gènes Hox. (Par ailleurs, le cluster Dd des téléostéens n'est pas largement représenté par l'ensemble des téléostéens). Cette figure a été réalisée à partir de plusieurs références [amores\\_zebrafish\\_1998](#); [guo\\_hox\\_2010](#); [lappin\\_hox\\_2006](#); [rux\\_hox\\_2017](#)

Ces duplications de génomes sont également étudiées pour la recherche médicale. Il a récemment été montré que dans plusieurs cas de tumeurs, des duplications de génome étaient impliquées (**gemble\_genetic\_2022**). Cette étude montre une forte instabilité génétique après des événements de WGD, ce qui favorisera les multiples mutations. Les duplications de génome sont également connues pour être suivies de perte massive de gènes dupliqués (**inoue\_rapid\_2015**; **jaillon\_genome\_2004**). Ce phénomène, qui est toutefois mal connu, s'appelle la rediploïdisation et consiste en un retour au nombre initial de gènes et non des chromosomes. Pour certains gènes, une seule copie des gènes sera maintenue, ce qui réduit grandement le nombre de paralogues issus de duplication de génome complet (**byrne\_yeast\_2005**). Par exemple, si le génome de la drosophile partage un patrimoine génétique en commun avec l'homme et compte environ 15 000 gènes codant des protéines, et avec les 2 duplications de génome qu'ont vécu les vertébrés, l'homme devrait avoir environ 60 000 gènes, or, il n'en compte qu'environ 22 000 gènes codants des protéines. Par ailleurs, certains gènes maintenus en duplicitat sont responsables de démence chez l'homme, comme la duplication du gène SNCA impliqué dans la maladie de Parkinson (**chartier-harlin\_alpha-synuclein\_2004**; **ibanez\_causal\_2004**). La question que nous pouvons nous poser est : quelle force évolutive régit les gènes restés en duplicitat ou revenus en singleton ?

### 1.1.3 Coévolution

Concernant la force évolutive qui agit sur les gènes, nous n'avons pas de réponse claire et universelle à donner. Cependant, des pistes pourraient expliquer cette pression, comme celle de la coévolution. La coévolution est le fait que des gènes évoluent ensemble en réponse les uns aux autres (**lovell\_integrated\_2010**). Et c'est ce que Thompson avait défini en 1994 comme étant la coévolution réciproque (**thompson\_coevolutionary\_1994**).

Un exemple assez parlant est celui des gènes du système immunitaire. Les gènes du système immunitaire n'ont pas d'autres choix que d'évoluer en fonction de l'évolution des pathogènes qui eux-mêmes évoluent en fonction du système immunitaire (**schlesinger\_coevolutionary\_**

La coévolution est également très étudiée pour les interactions ligands récepteurs, en effet cette interaction est soumise à une certaine pression notamment au niveau de leur site de liaison qui peut être unique en fonction de la spécificité de la relation. Pour le couple OXT-OXTR (l'ocytocine et son récepteur) par exemple, les séquences moléculaires ont longtemps été pensées comme intactes chez les mammifères, seulement, il a été montré quelques différences subtiles qui seraient corrélées avec les changements comportementaux sexuels chez les primates étudiés (**vargas-pinilla\_evolutionary\_2015**).

### 1.1.4 Naissance et mort des gènes

Entre les mutations ponctuelles, les duplications de génome, et les pertes massives qui ont suivi, il est question de la naissance et de la mort des gènes. La revue de **kaessmann \_ origins \_ 2010** détaille les différents types de naissance d'un gène. Nous allons en expliquer quelques-uns.

Dans un premier temps, un nouveau variant d'un gène peut naître d'une mutation ponctuelle. En effet, une mutation peut modifier la séquence nucléotidique d'un gène déjà existant sans qu'il soit pseudogénisé. Un nouveau gène peut apparaître et potentiellement une nouvelle protéine fonctionnelle.

Dans un deuxième temps, la duplication est un moteur clé de la création de nouveaux gènes, que la duplication soit une duplication complète du génome, d'un chromosome uniquement ou même juste localement d'un gène ou d'une partie de séquence.

La naissance des gènes est importante pour la spéciation et l'adaptation des espèces. Par ailleurs, ce sont des processus très longs, tout comme la mort d'un gène. Si un gène peut naître d'une duplication ou d'une mutation, ce sont les mêmes processus qui peuvent l'amener à la mort. Comme nous l'avons vu dans le Chapitre ?? page ??), une mutation impliquant un codon stop prématûr peut aboutir à un pseudogène.

Nous pouvons déterminer si un gène est mort en comparant sa séquence à celle d'une autre espèce proche, mais pour laquelle le gène est fonctionnel. À partir de l'alignement de séquence, il est possible de retrouver, comme dans le cas de figure présenté en Figure ??, une mutation qui remplace un acide aminé par un codon stop (\*) par exemple.

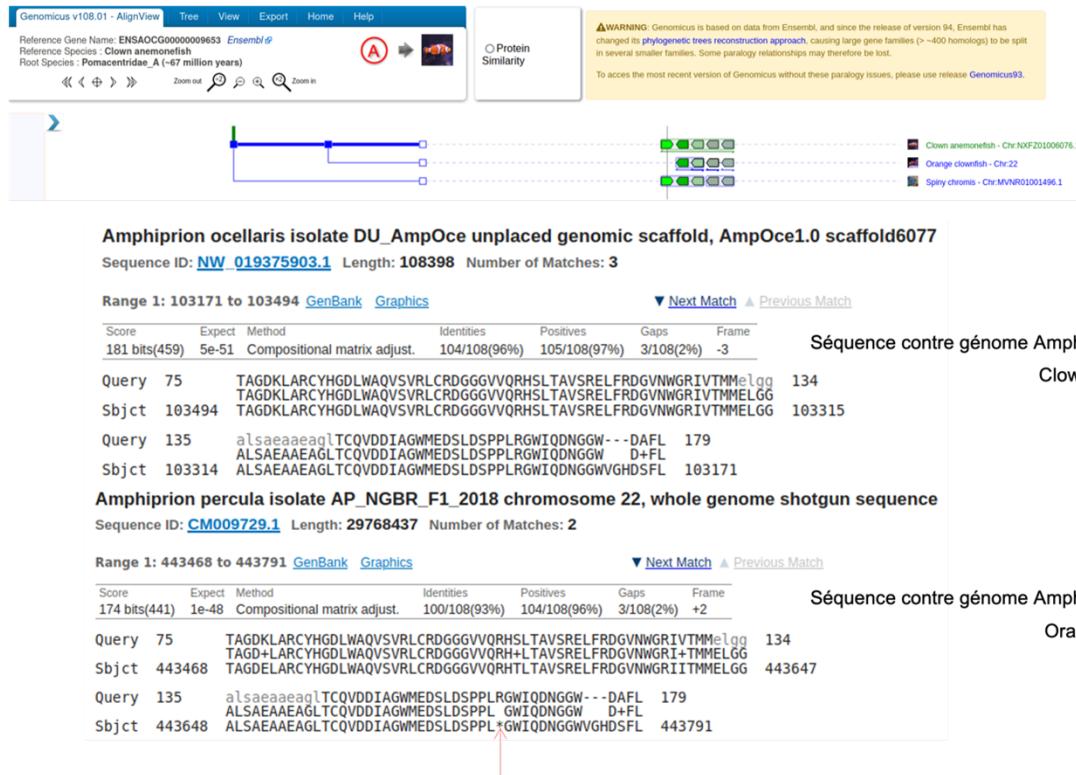


FIGURE 1.4 – Exemple d'un pseudogène avec BLAST

Légende : Dans la première partie de la figure, il s'agit d'une capture d'écran de l'alignement des gènes (synténie, décrit en chapitre ??), centré sur le gène en vert (ENSAOCG00000009653, BCL2) qui est présent chez les espèces *Clown anemonefish* et *Spiny chromis*, mais absent chez *Orange clownfish*. Il y a donc une suspicion de pseudogène, car la séquence semble être conservée pour les gènes à proximité. Une fois la séquence protéique du gène bcl2 de l'espèce *Clown anemonefish* récupérée, on l'aligne sur son génome. Aucune particularité à observer. Puis, on l'aligne sur l'espèce *Orange clownfish*, et là, on remarque qu'il y a un codon stop (\*) qui s'est glissé dans la séquence au niveau de l'Arginine. Le gène est bien pseudogénisé chez cette espèce.

## 1.2 Phylogénie des animaux

La phylogénie s'est définie graduellement. D'abord avec le Suédois Carl Von Linné au XVIII<sup>e</sup> siècle qui s'impose avec l'idée de biodiversité et sa classification de plus de 10 000 espèces en binôme genre-espèce. Ce sont ensuite les travaux de Lamark au XIX<sup>e</sup> siècle qui évoquent les liens de parenté et la transmission du patrimoine génétique d'un parent à ses descendants. Et c'est en 1859 que Charles Darwin publie son ouvrage « De l'origine des espèces » qui soudera la théorie de l'évolution selon laquelle les êtres vivants sont issus d'ancêtres communs, et l'évolution des espèces se produit grâce à la sélection naturelle (caractère héréditaire favorisant la reproduction et la survie de l'espèce).

Avec le temps, les méthodes se sont vues améliorées, nous sommes passés d'une classification morphologique à une classification par la séquence ADN.

### 1.2.1 Principe de la phylogénie

Le principe de phylogénie reprend notamment les travaux de Charles Darwin, et repose principalement sur le fait que chaque organisme descend d'un ancêtre commun. On représente communément les relations de parenté entre les organismes dans un arbre. Les arbres sont constitués de branches qui représentent les liens de parentés, et les nœuds les points de divergence. Ils sont construits à partir d'alignement de séquences et permettent de refléter la proximité entre celles-ci.

Différentes méthodes existent afin de parvenir à représenter l'histoire évolutive des séquences :

- Méthode de parcimonie (la méthode la plus simple, car on considère le minimum d'événement évolutif),
- Méthode de distance (on considère la distance entre les espèces par le nombre de différences entre les séquences),
- Méthode basée sur le maximum de vraisemblance (méthode qui va utiliser les probabilités d'obtention d'un arbre en fonction de plusieurs scénarios),
- ou encore des méthodes plus complexes pour des arbres comportant des hybridations et donc ne peut être définie de manière arborescente (**tagu\_bio-informatique\_2010**).

Chaque méthode de reconstruction d'arbre a ses avantages et ses faiblesses, et dépend principalement des données et des objectifs. Cependant, il est préférable de combiner plusieurs méthodes et d'utiliser une méthode d'estimation de la robustesse des nœuds une fois l'arbre généré. La robustesse est estimée par la méthode de *bootstrap* par exemple qui consiste à compter le nombre de fois qu'une branche est présente dans un pool d'arbres échantillons obtenu à partir d'alignements pris aléatoirement dans notre jeu de données.

Deux types d'études peuvent être considérés en phylogénie. Dans un premier temps, il y a la phylogénie des espèces, qui place des espèces par rapport à d'autres. Et dans

un deuxième temps, il y a la phylogénie des gènes qui permet de mettre en relation des similitudes des gènes. Les arbres de gènes peuvent être différents des arbres des espèces. Pour exemple, le gène FOXP2, aussi appelé le « gène de la parole » est présent chez des espèces vertébrées comme l'homme, la souris, la chauve-souris ou encore des oiseaux ([enard\\_molecular\\_2002](#); [scharff\\_evolutionary\\_2005](#); [webb\\_foxp2\\_2005](#); [white\\_singing\\_2007](#)). La séquence protéique est fortement conservée chez toutes les espèces, à l'exception de la chauve-souris qui pourrait être liée à l'écholocalisation chez celle-ci ([li\\_accelerated\\_2007](#)). En représentant l'arbre du gène FOXP2 et l'arbre des espèces, on se rend compte qu'ils ne sont pas identiques ou congruents (Figure ??).

Dans notre cas, nous avons travaillé avec des arbres phylogénétiques de gènes et certains concepts et termes sont à définir, comme l'homologie, l'orthologie et la paralogie.

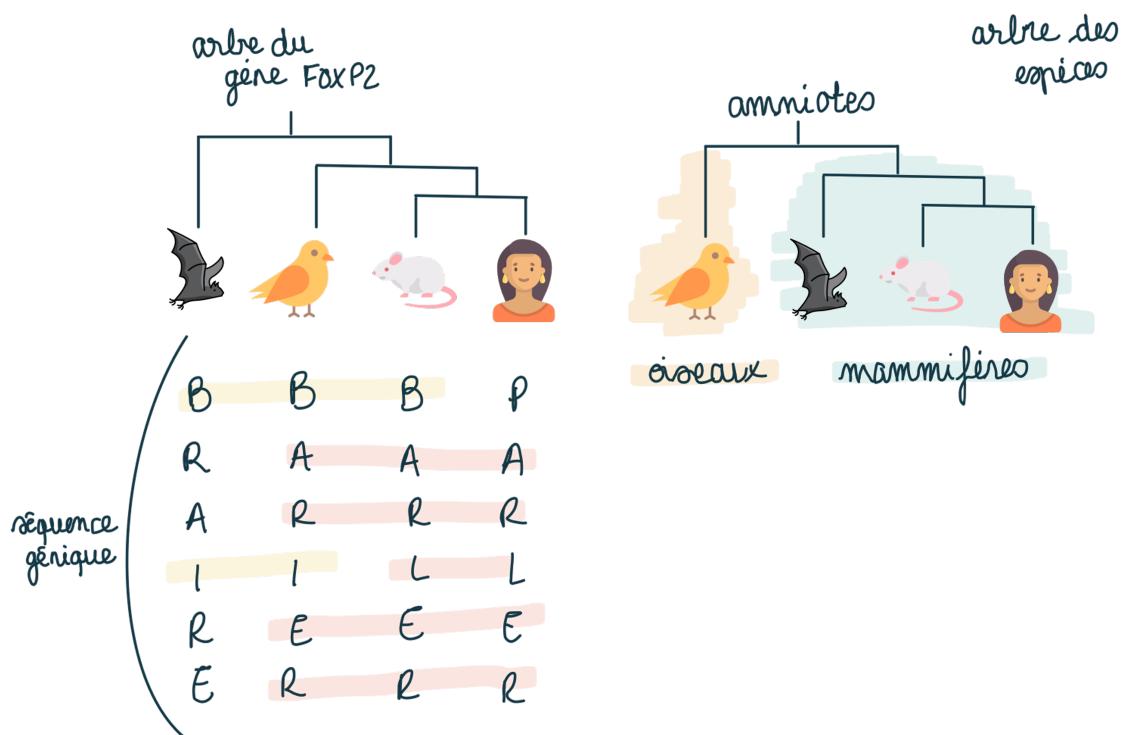


FIGURE 1.5 – Différences entre arbre de gènes et d'espèces

Légende : À gauche, il y a l'arbre du gène FOXP2 en fonction de l'alignement des différentes séquences. À droite, il y a l'arbre des espèces.

### 1.2.2 Homologie, orthologie et paralogie

Homologie, orthologie et paralogie désignent tous les 3 des degrés différents d'évolution de parcours des gènes.

Le terme homologie se réfère aux similitudes de séquences entre plusieurs gènes, il englobe les termes paralogie et orthologie. Deux gènes ayant un ancêtre commun sont dits « homologues ».

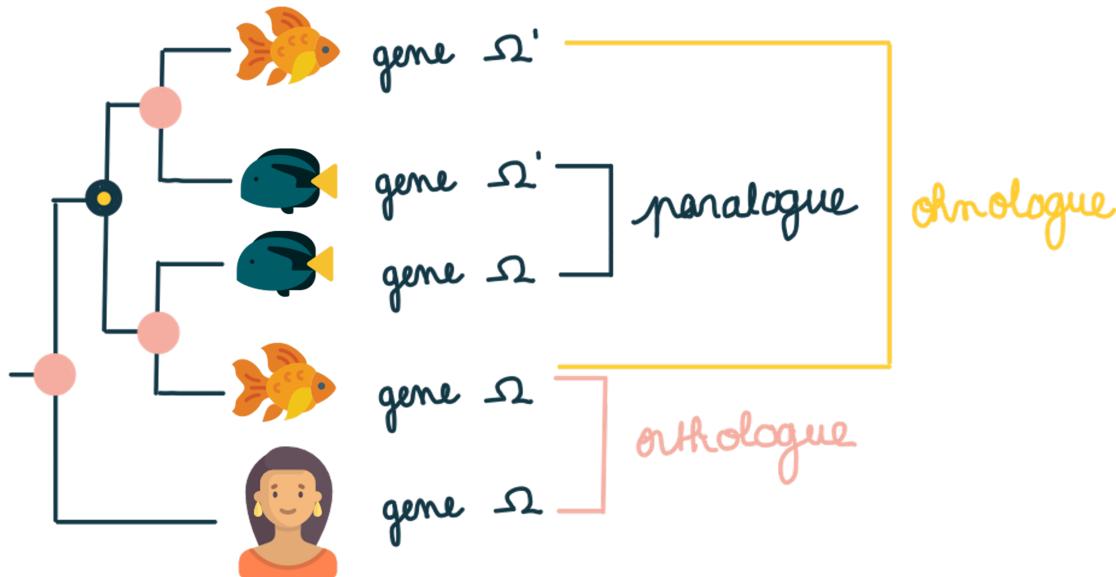
La paralogie fait référence à des gènes issus d'une même descendance, mais ayant subi des duplications génétiques au sein d'une même espèce. Ils peuvent avoir subi également des mutations qui ne les rendent pas identiques. De plus, les gènes paralogues n'ont pas forcément la même fonction, il est même plutôt fréquent d'avoir une néo-fonctionnalisation (nouvelle fonction du gène), une sous-fonctionnalisation (les deux gènes acquièrent une sous-fonction de la fonction mère du gène), une spécialisation spatio-temporelle ou une spécialisation de fonction (**kuzmin\_retention\_2022**).

L'orthologie concerne les gènes partageant un ancêtre commun suivi d'un événement de spéciation, donc d'espèces différentes (**fitch\_distinguishing\_1970**). La définition d'un orthologue est simple lorsqu'on parle de deux gènes, mais se complique rapidement lorsqu'on parle de plusieurs gènes pour plusieurs espèces. Et pourtant, c'est bien grâce aux relations d'orthologie que l'on a fait des avancées en génomique comparative et pour l'annotation des nouveaux génomes séquencés (**huerta-cepas\_fast\_2017**).

Les relations d'homologies sont donc étroitement liées les unes avec les autres et peuvent être complexes. Au sein des arbres phylogénétiques, il y a à la fois, les orthologues et les paralogues. Pour être plus précis, lorsque des orthologues n'ont pas subi de duplication par la suite, ils sont appelés orthologues directs. Le terme parologue direct existe également pour les gènes dupliqués au sein d'une même espèce. (Figure ??).

Les gènes paralogues issus de duplication de génome sont appelés onhologue en mémoire à Ohno qui a mis en lumière les duplications de génome chez les vertébrés (**ohno\_evolution\_1963**). Les gènes ohnologues représentent entre 20 et 35% des gènes humains et concernent majoritairement les gènes de transduction du signal (74%) (**singh\_identification\_2015**).

De plus, il existe également des catégories pour parler du nombre d'orthologues du gène dans deux espèces. Premièrement, nous avons l'orthologie un à un (1 :1), qui correspond à un orthologue strict chez une espèce d'un gène chez une autre espèce. Ensuite il y a les co-orthologues qui peuvent être un-à-plusieurs ou plusieurs-à-un (1 :n), qui comme son nom l'indique évoque un exemplaire d'un gène chez une espèce, et plusieurs dans une autre. Puis les orthologues plusieurs-à-plusieurs (m ;n), ainsi que les orthologies complexes prenant en compte les pertes de gènes, les duplications de génomes et autres événements d'évolutions complexes (Figure ??) (**koonin\_orthologs\_2005**).



- spéciation
- duplication
- duplication complète des génomes

FIGURE 1.6 – Représentation des différents homologues

Légende : En rose sont représentés les orthologues à la suite d'une spéciation. En noir, les paralogues à la suite d'un évènement de duplication et en jaune les gènes ohnologues à la suite d'une duplication complète des génomes.

### 1.2.3 Arbre de la vie des animaux

Comme évoqué dans le chapitre précédent, les relations entre les espèces peuvent être représentées en arbre, appelé « arbre des espèces ». Ces arbres renferment la proximité des espèces (proximité des branches), leur temps de divergence (longueur des branches) ainsi les évènements vécus au fil du temps (nœud) avec comme exemple les duplications de génome entre autres. Les arbres phylogénétiques utilisent également des termes de la classification taxonomique des espèces. Mais, il existe des biais à l'élaboration d'un arbre de la vie comme la convergence évolutive. La convergence évolutive est le mécanisme par lequel des espèces n'ayant pas d'ancêtre commun proche évoluent dans la même direction. En d'autres termes, elles n'ont pas reçu de caractère commun de leur descendance commune, mais l'ont acquis indépendamment l'une de l'autre. En génomique, la convergence évolutive se traduit par une mutation similaire dans des gènes spécifiques distincts. Ça peut être le cas notamment pour les résistances aux bactéries. Cela peut

donc devenir un biais dans le sens où des évènements de mutations semblables ayant lieu après divergence des espèces peuvent rapprocher ces deux espèces dans l'arbre de la vie par erreur (**christin causes 2010**). Il a été montré que les systèmes olfactifs des vertébrés et celui des protostomiens partagent des arrangements physiologique similaires (**hildebrand mechanisms 1997**) mais diverses origines (**strausfeld olfactory 1999**).

Dans notre cas, nous utilisons à tort le terme « arbre de la vie des animaux » car nous étudions l'ensemble du règne animal ainsi qu'une espèce du règne des champignons. L'ensemble des deux règnes est nommé Opistoconte *Opisthokonta*) (Figure ??).

Dans notre étude, nous avons dû partir de l'homme, car c'est l'espèce la plus référencée, et la plus étudiée au niveau des voies de signalisation, comme nous le verrons par la suite. De ce fait, l'utilisation du terme « remonter l'arbre » sera utilisé à l'avenir car on remonte les branches de l'arbre à partir de l'homme pour aller vers ses ancêtres communs avec d'autres espèces sur le chemin.

Plus nous parcourons l'arbre vers l'ancêtre commun le plus éloigné, et moins nous aurons de similitude avec les différentes espèces. Au sein de l'espèce, nous partageons notre génome à 99,9%, c'est ce 0,01% qui explique nos différences comme la couleur de nos cheveux, la couleur de nos yeux, notre taille ou nos allergies. Nous partageons 98,7% de notre patrimoine génétique avec le chimpanzé, 90% avec le chat domestique, 85% avec la souris, et plus étonnant, 26% avec la levure et 18% avec les champignons de Paris (**roy biotechnology 2010**) Ces valeurs sont déterminées en utilisant uniquement les gènes codants. Or, les gènes codants ne représentent que 5% du génome humain. Et par complémentarité, le génome humain est constitué à 6% de gènes communs à l'ensemble des primates, 13% aux vertébrés, 16% aux animaux, 28% aux eucaryotes et 37% aux bactéries (**domazet-loso ancient 2008 ; mcfall-ngai animals 2013**).

De connaître ces similitudes a permis de connaître et comprendre l'arbre de la vie dans son entiereté. C'est tout particulièrement important dans notre étude, car nous allons remonter l'arbre de la vie des Opistoconte (Figure ??), pour connaitre l'origine des gènes impliqués dans une voie de signalisation. Cet arbre simplifié des différents nœuds et clade est notre point de repère pour la première étude. Un nombre de 25 clades a été sélectionné pour définir nos moments d'apparitions. L'estimation du moment où un gène est apparu se base sur l'observation de son orthologue le plus éloigné dans l'arbre phylogénétique, c'est-à-dire le point ancestral le plus éloigné où un orthologue du gène humain peut être identifié. Pour illustrer, si un gène humain possède un orthologue chez les poissons, mais que cette homologie ne se poursuit pas au-delà, on peut supposer que le moment de l'apparition de ce gène remonte probablement au nœud des *Eutelostomi*.

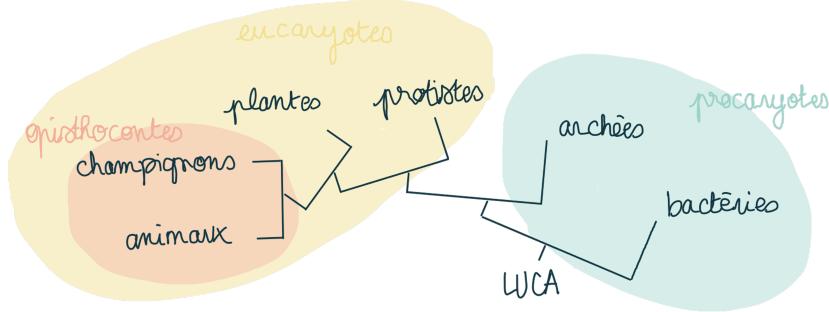


FIGURE 1.7 – Arbre de la vie

Légende : LUCA pour Last Universal Common Ancestor, en français le Dernier ancêtre commun universel.

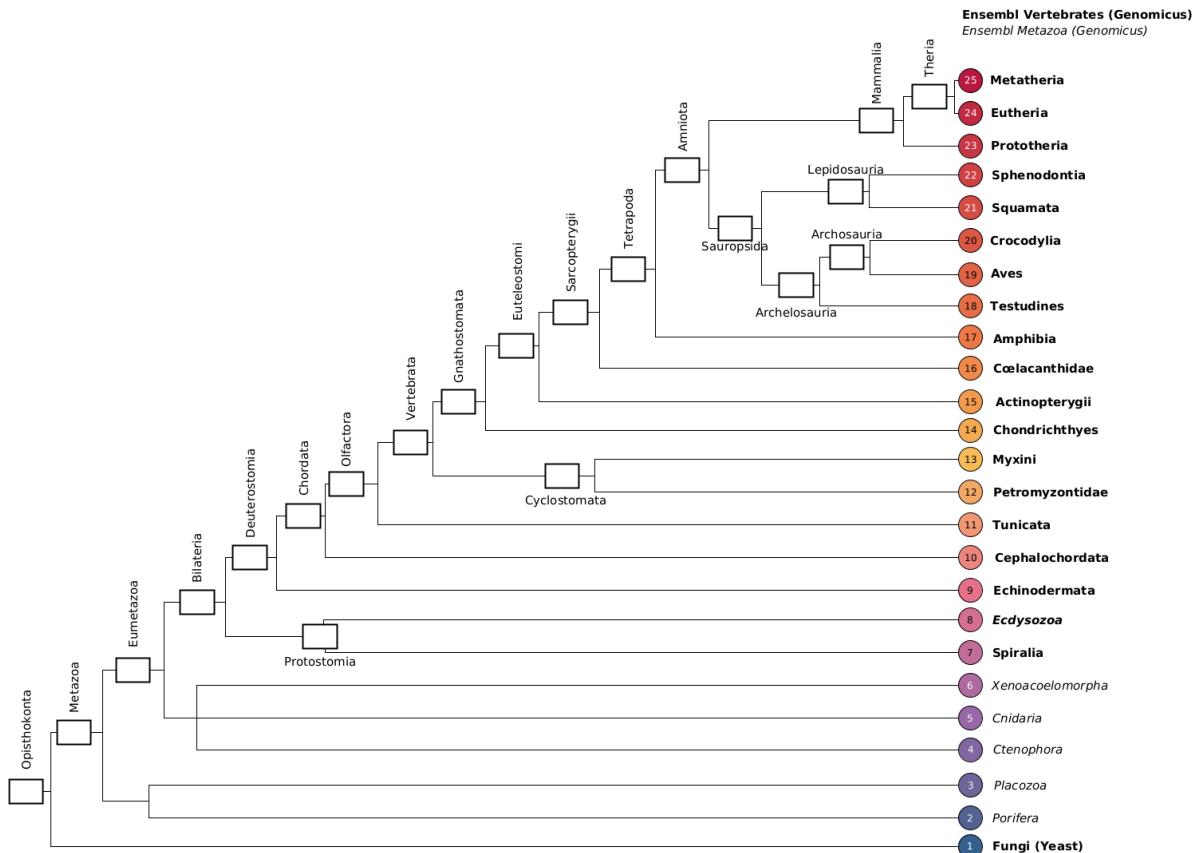


FIGURE 1.8 – Arbre de la vie des Opisthokontes simplifié pour notre étude

Légende : Chaque rectangle représente un nœud de spéciation au cours de l'évolution. Chaque branche représente un clade terminal de l'arbre avec son numéro associé dans les ronds de couleurs. La longueur des branches n'est pas représentative de la divergence évolutive réelle. En gras sont les clades disponibles sur Genomicus Vertebrates, et en italique les clades disponibles sur Genomicus Metazoa.

### 1.2.4 Spécificité des téléostéens

Un clade parmi les animaux va nous intéresser particulièrement pour l'étude 2, il s'agit des poissons téléostéens (*Teleostei*). L'infra-classe des téléostéens fait partie du sous-embranchement des vertébrés, qu'ils représentent largement car approximativement 50% des vertébrés sont des téléostéens. Environ 25 000 espèces composent ce clade, et la diversité de ce groupe permet difficilement une définition globale. Cependant, ils se caractérisent par leur ossature complète, ils ont des nageoires rayonnées, et une mâchoire. Ce qui exclut les poissons cartilagineux comme les raies et les requins. D'un point de vue morphologique et écologique, les téléostéens sont très diversifiés. Certains peuvent vivre en eaux douces ou en eaux salées, en eaux profondes et pour quelques-uns avoir des interactions avec le monde terrestre. Ils peuvent également avoir des stratégies de reproduction variées (ovipare, vivipare et même ovovivipare).

Cette diversité pourrait être due à la troisième duplication de génome complet partagée par l'ensemble des espèces de téléostéens en plus des deux survenues pour l'ensemble des vertébrés (**ravi\_rapidly\_2008**; **taylor\_genome\_2003**). Par ailleurs, les téléostéens ne dérogent pas à la règle et leur troisième duplication de génome est également suivie d'une période de rediploïdisation (perte massive de gènes dupliqués) (**inoue\_rapid\_2015**).

Une troisième duplication est donc spécifique aux téléostéens, mais à cela s'ajoute une quatrième duplication pour deux clades distincts des téléostéens : les salmonidés et les carpes (**jaillon\_genome\_2004**; **lien\_atlantic\_2016**) (Figure ??).

La découverte de cette troisième duplication de gènes a pour origine l'identification de 7 complexes Hox en 1998 chez le poisson zèbre contre 4 chez l'homme (**amores\_zebrafish\_1998**) tandis qu'on en trouvera 8 chez le *Japanese eel* qui est également un téléostéen (**guo\_hox\_2010**) (Figure ??).

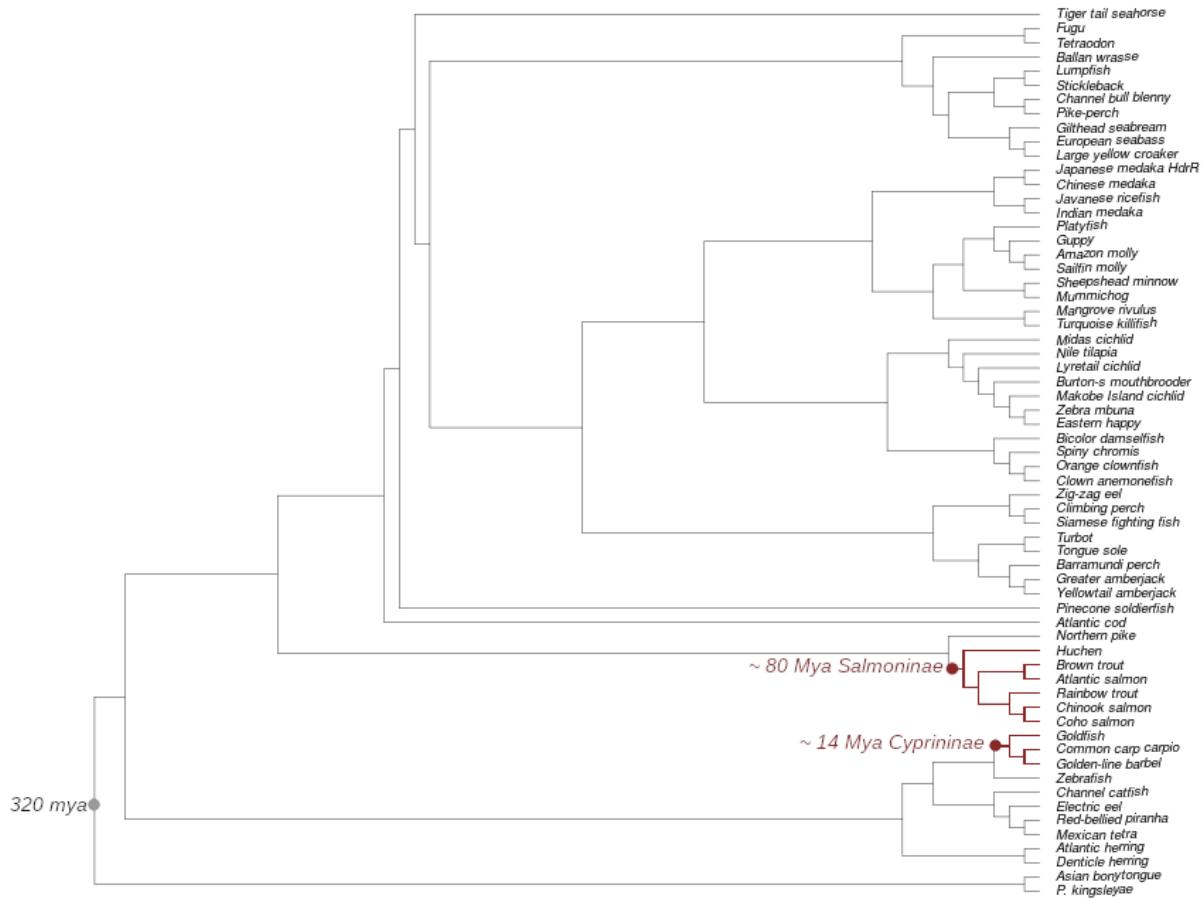


FIGURE 1.9 – Arbre des téléostéens

Légende : Les nœuds avec un point correspondent au moment de duplication de génome. Les branches noires sont les espèces 3 WGD, et en rouge les espèces 4 WGD.

### 1.2.5 Bases de données

Durant cette étude, nous avons utilisé différentes bases de données mettant à disposition des arbres phylogénétiques, comme notamment la base de données Ensembl, ainsi que la base de données Genomicus. Genomicus est basée sur les données d'Ensembl, et les deux bases de données existent pour différents règnes : *Vertebrates*, *Bacteria*, *Fungi*, *Plants*, *Protistes* et *Metazoa*. Dans les deux sous-chapitres suivants, nous décrirerons le fonctionnement général pour Vertebrates, mais ils fonctionnent tous de la même façon.

#### Ensembl

Premièrement, il est important de présenter la base de données Ensembl (<https://www.ensembl.org/>). Il s'agit d'une base de données regroupant nombreuses ressources génomiques. Elle offre une annotation complète des gènes, permet de réaliser des alignements de séquences et calcule les multiples alignements. Nous avons utilisé différentes versions

de la base de données. Chaque version étant la plus récente au moment des études. La dernière version que nous avons utilisée est la version V109. Elle comprend 199 espèces de vertébrés. Dans notre cas, nous avons utilisé Ensembl BioMart notamment pour l'étude 2. BioMart est un outil d'Ensembl permettant des requêtes complexes et automatisées au sein de la base de données comme la recherche d'orthologue par exemple, ou de paralogues, de pourcentages d'homologie (Figure ??).

Dans notre cas, nous l'avons utilisé afin de récupérer l'ensemble des orthologues téléostéens des gènes humains codant des protéines. En version 107, il y avait 63 espèces de téléostéens, dont 54 espèces 3 WGD et 9 espèces 4 WGD (Figure ??).

## Genomicus

Genomicus (<https://www.genomicus.bio.ens.psl.eu/>) est une base de données « fille » de la base de données Ensembl car elle utilise les arbres Ensembl pour ses propres outils. Cependant, les arbres Ensembl sont modifiés pour résoudre les problèmes de noeuds de duplications mal documentées. La structure de l'arbre est alors révisée en transformant ces nœuds de duplications mal établis en nœuds de spéciation, tout en réaffectant les quelques gènes dupliqués vers des nœuds plus récents (**louis\_genomicus\_2015**).

Ces arbres sont accessibles au format .nhx (New Hampshire eXtended, format basé sur New Hampshire eXtended, format standard pour les arbres phylogénétique).

Genomicus propose également une visualisation de la synténie des gènes sur les chromosomes, ce qui aura des avantages lors de la recherche de pseudogènes. Le terme synténie vient de la conservation plus ou moins grande de l'ordre des gènes sur un chromosome, relativement entre eux au sein des génomes. (Figure ??). La notion de synténie découle de l'étude comparative des génomes. C'est notamment grâce à cette synténie que nous pouvons différencier les duplications complètes de génome des duplications ponctuelles comme ça a été l'exemple avec la famille des gènes Hox (Figure ??).

**Dataset**: Human genes (GRCh38.p13)

**Filters**: Gene type: protein\_coding

**Attributes**: Gene stable ID, Gene name, Zebrafish gene stable ID, Zebrafish gene name

Gene stable ID	Gene name	Zebrafish gene stable ID	Zebrafish gene name
ENSG00000198888	MT-ND1	ENSDARG00000063895	mt-nd1
ENSG00000198763	MT-ND2	ENSDARG00000063899	mt-nd2
ENSG00000198804	MT-CO1	ENSDARG00000063905	mt-co1
ENSG00000198712	MT-CO2	ENSDARG00000063908	mt-co2
ENSG00000228253	MT-ATP8	ENSDARG00000063911	mt-atp8
ENSG00000198899	MT-ATP6	ENSDARG00000063912	mt-atp6
ENSG00000198938	MT-CO3	ENSDARG00000063914	mt-co3
ENSG00000198840	MT-ND3	ENSDARG00000063914	mt-nd3
ENSG00000212907	MT-ND4L		
ENSG00000198886	MT-ND4	ENSDARG00000063917	mt-nd4

FIGURE 1.10 – Affichage des résultats Ensembl Biomart

Légende : Dans cet exemple, nous avons effectué la recherche sur la version Ensembl Biomart V107 des orthologues *Zebrafish* de gènes codant des protéines humaines.

Genomicus v108.01 - AlignView Tree View Export Home Help

Reference Gene Name: MT-ND1 (ENSG00000198888) Ensembl

Reference Species: Human  
Root Species: Fungi/Metazoa group (~1500 million years)

(A) →

○ Protein Similarity

⚠ WARNING: Genomicus is based on data from Ensembl, and since the release of version 94, Ensembl has changed its phylogenetic trees reconstruction approach, causing large gene families (> ~400 homologs) to be split in several smaller families. Some paralogy relationships may therefore be lost.

To access the most recent version of Genomicus without these warnings

A more recent version of genomicus is available here

Human - Chr:MT      Zebrafish - Chr:MT

FIGURE 1.11 – Affichage des résultats Genomicus

Légende : À partir des résultats Ensembl Biomart, on a été regarder la synténie du gène MT-ND1 entre l'Homme et le *Zebrafish*. Le gène MT-ND1 est coloré en vert. On peut voir que la synténie de cette zone est conservée à l'exception de deux gènes perdus chez *Zebrafish*.

## 1.3 Communication cellulaire

Les voies de signalisation sont un moyen de transmission d'informations pour et par les cellules. Par ce biais, elles permettent le développement, la croissance, le maintien homéostatique des cellules d'organismes multicellulaires (**combarinous\_communications\_2013**). Les voies de signalisations se caractérisent par une cascade d'interactions protéiques au sein d'une cellule (Figure ??). Cette cascade est initiée par un ligand se fixant à son récepteur ou en réponse à un *stimulus*, et se terminant par une réponse cellulaire ou la régulation d'un gène. Il existe plusieurs grandes catégories de voie de signalisation qui sont elles-mêmes caractérisées par le type du récepteur.

L'étude des voies de signalisation est nécessaire, notamment pour en apprendre plus sur le fonctionnement des maladies ou les dysfonctionnements entre interactions protéiques. Les études sur les interactions au sein des cellules s'appellent l'interactome (ensemble des interactions d'une cellule).

### 1.3.1 Interactions protéiques

Lorsqu'on parle d'interactions protéiques, on peut rapidement parler de biologie systémique car elle fait écho aux études de réseaux complexes d'un organisme. Les protéines interagissent entre elles, mais ne font généralement pas intervenir toute la molécule. Les interactions se font via des sites de liaison. Les protéines peuvent donc se lier soit spécifiquement avec une protéine partenaire unique, mais peuvent également se lier avec plusieurs partenaires. Ces spécificités d'interactions peuvent être des interactions spécifiques à différentes échelles, au sein d'une cellule, au sein d'un organisme, et même plus largement commune à plusieurs organismes. Dans le cas des interactions multiples, on parle alors d'un site de liaison qui est peu spécifique, et/ou d'une protéine avec plusieurs sites de liaisons (**di\_lullo\_mapping\_2002**). Les interactions peuvent être directes, mais peuvent également se faire à distance grâce à des forces permettant la liaison entre acides aminés :

- Les liaisons électrostatiques et ioniques qui se produisent entre charges opposées,
- Les liaisons hydrophobes qui se produisent lorsque les régions hydrophobes des protéines s'attirent,
- Les liaisons hydrogènes se forment lorsque l'atome d'hydrogène est partagé entre deux atomes, l'un étant plus électronégatif que l'autre (**bondar\_hydrogen\_2012**),
- Les liaisons par ponts hydrogènes entre protéines et molécules,
- Les liaisons de Van Der Waaals qui se produisent entre atomes non polaires (**van\_oss\_role\_1986**)

Toutes ces formes différentes de liaisons permettent un panel d'interactions possibles entre les protéines. En multipliant les possibilités d'interactions, on peut s'imaginer qu'elles n'ont pas besoin d'être maintenues dans le temps puisqu'elles trouveront d'une façon ou

d'une autre un partenaire avec qui interagir au sein d'une espèce. En 2006, une étude a montré que parmi 70 000 interactions protéine-protéine, seulement 16 étaient communes aux 4 espèces étudiées (vers, mouche, levure et homme) ([gandhi\\_analysis\\_2006](#)).

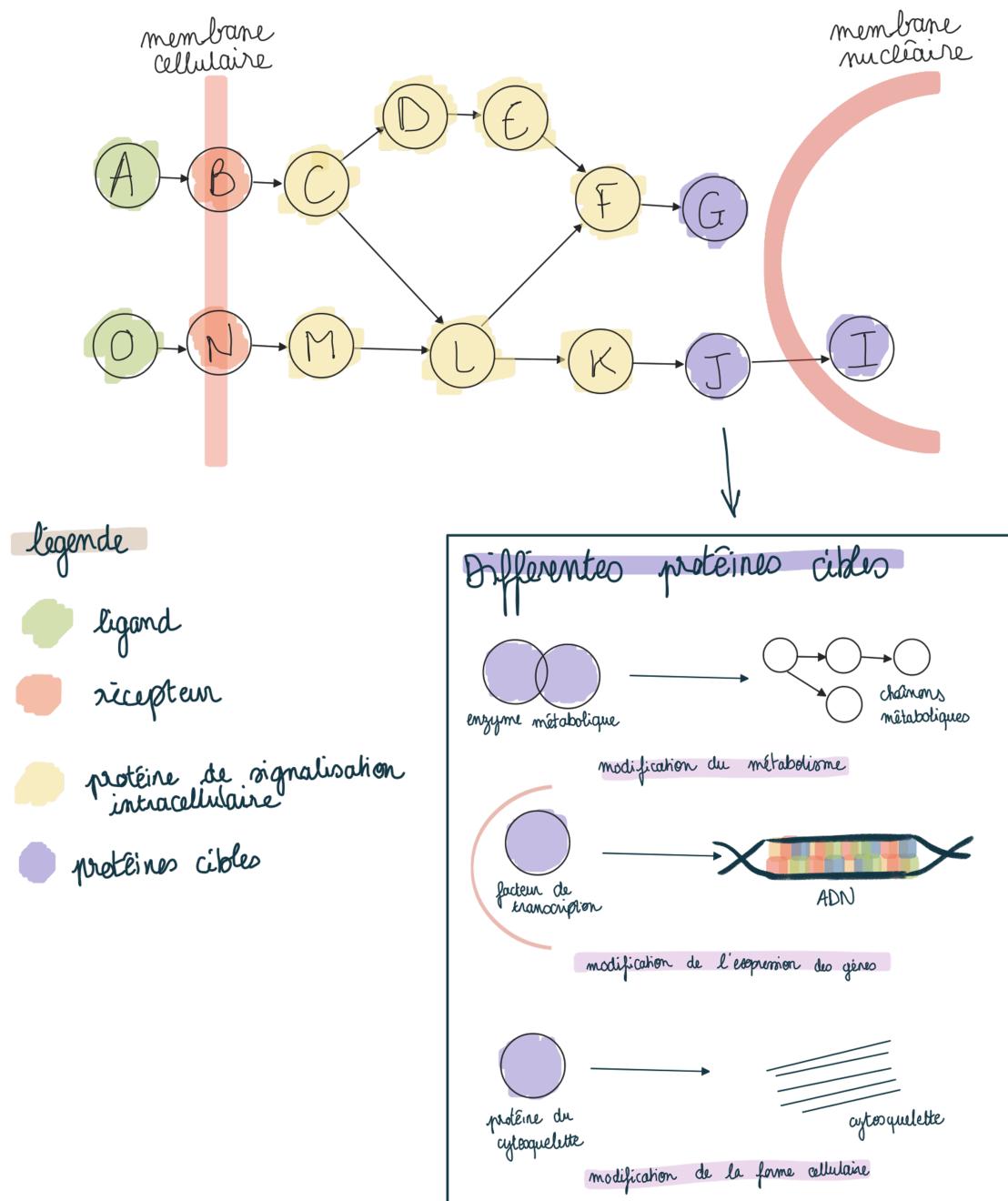


FIGURE 1.12 – Schéma d'une voie de signalisation intracellulaire

Légende : Chaque rond représente une protéine et les flèches les interactions.

### 1.3.2 Relation ligand-récepteur

Les ligands et les récepteurs membranaires sont deux protéines interagissant ensemble au niveau de la membrane d'une cellule. Le ligand se fixant à son récepteur va entraîner l'activation des molécules de la voie de signalisation intracellulaire (Figure ??).

Les relations ligand-récepteur peuvent être, comme toute interaction, soit spécifique, soit non spécifique. Par exemple, la voie JAK-STAT a environ 60 ligands et 40 récepteurs connus (**darnell\_jak-stat\_1994**). À l'inverse, la voie de l'insuline est très spécifique de 3 ligands et 2 récepteurs connus (**leroith\_insulin-like\_2021**). Pour vulgariser les relations ligand-récepteur, on parle très souvent de clé (ligand) et serrure (récepteur). Une clé peut ouvrir plusieurs serrures, et une serrure peut être ouverte par plusieurs clés, mais les deux ont généralement besoin l'un de l'autre pour fonctionner.

De nombreuses études portent sur la coévolution des interactions ligand-récepteur notamment pour des questions médicales, notamment car les médicaments ciblent généralement les récepteurs (**de\_jong\_receptor-ligand\_2005** ; **pluder\_proteome\_2006** ; **woolhouse\_biological\_2002**).

### 1.3.3 KEGG

Concernant les interactions et les voies de signalisation en général, il a fallu trouver une source de données nous permettant d'avoir accès à ces informations. Et la base de données KEGG (Kyoto Encyclopedia of Genes and Genome) (<https://www.genome.jp/kegg>) est l'une des premières à représenter exhaustivement les interactions au sein d'une cellule sous forme de voies (**bader\_pathguide\_2006**) et une des plus complètes avec 568 voies de signalisation décrite manuellement dont 335 voies humaines (**chanumolu\_kegg2net\_2021**).

La base de données a plusieurs fonctions comme la visualisation d'une multitude de voies, mais également la coloration des gènes au sein des voies, et ça sera très utile dans notre cas. Sur KEGG, les voies sont représentées chez l'homme, et quelques fois pour la levure, la drosophile ou encore plusieurs espèces à la fois.

Elle permet également de récupérer les voies dans un format utilisables par des bibliothèques R. Chaque voie de signalisation a donc été récupérée au format .xml (Extensible Markup Language) à partir de l'outil de base de données PATHWAY de KEGG. Ce format de fichier est bien pris en charge par les bibliothèques R telles que XML (**lang\_xml\_2023**) et igraph (**csardi\_igraph\_2023** ; **csardi\_igraph\_2006**) (Figure ??).

Pour les deux études, nous avons utilisé la base de données KEGG V104.0. À partir des mots de clés « *signaling pathway* » et « *human* », nous avons récupéré un ensemble de 47 voies de signalisation représentant 2 298 gènes uniques. L'ensemble des voies sont représentées ainsi que leurs caractéristiques dans le tableau ??.

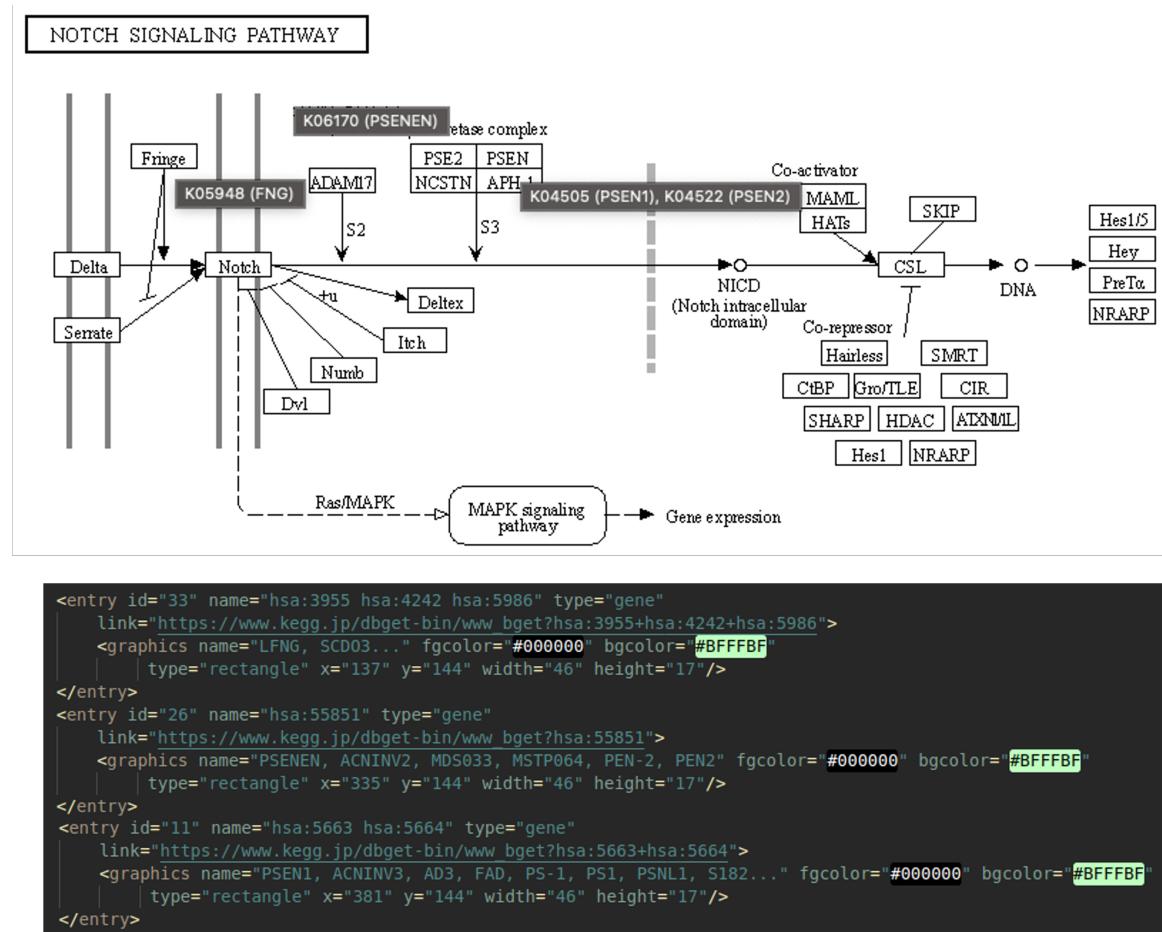


FIGURE 1.13 – Représentation d'une voie KEGG

Légende : A. La représentation graphique d'une voie de signalisation KEGG sur la page internet avec l'exemple de la voie Notch. Chaque rectangle représente un gène et les flèches une interaction ( $\rightarrow$  activation,  $\dashv$  inhibition). Les protéines qui se collent sont des complexes protéiques. B. Un extrait de cette même voie en format fichier .xml. Chaque gène est représenté par une balise « *entry* » comprenant un identifiant, un nom et un type.

TABLE 1.1 – Caractéristiques des voies de signalisation étudiées

Nom de la voie	Catégorie KEGG	Nombre de gènes dans la voie	Nombre d'interactions gene-gene	Nombre de sous-voie	Nombre de gènes dans la plus grande sous-voie	Etendue des moments des naissances des gènes
p53	Cell growth and death	64	70	254	7	[1,25]
AGE-RAGE	Endocrine and metabolic disease	62	93	358	6	[1,24]
Adipocytokine	Endocrine system	36	48	54	8	[1,17]
Estrogen	Endocrine system	62	69	29	16	[2,24]
Glucagon	Endocrine system	50	55	40	9	[1,25]
GnRH	Endocrine system	41	39	16	16	[1,21]
Insulin	Endocrine system	62	77	255	11	[1,25]
Ovarian steroidogenesis	Endocrine system	45	25	15	6	[1,25]
Oxytocin	Endocrine system	58	73	51	9	[1,21]
PPAR	Endocrine system	51	63	61	2	[1,23]
Prolactin	Endocrine system	54	55	63	14	[1,25]
Relaxin	Endocrine system	81	103	140	13	[1,24]
Thyroid hormone	Endocrine system	78	85	102	7	[1,21]
B cell receptor	Immune system	47	57	88	13	[1,22]
C-type lectin receptor	Immune system	154	185	107	8	[1,25]
Chemokine	Immune system	58	82	183	11	[1,21]
FC epsilon RI	Immune system	42	47	42	10	[1,24]
IL-17	Immune system	91	152	709	9	[1,25]
NOD-like receptor	Immune system	168	164	420	9	[1,25]
RIG-I-like receptor	Immune system	53	73	374	8	[1,19]
T cell receptor	Immune system	66	99	376	9	[1,25]
Toll-like receptor	Immune system	79	109	371	10	[1,25]
Neurotrophin	Nervous system	77	117	952	16	[1,25]
AMPK	Signal transduction	69	67	184	10	[1,25]
Apelin	Signal transduction	62	78	26	11	[1,21]
Calcium	Signal transduction	52	67	77	8	[1,21]
cAMP	Signal transduction	88	120	1966	11	[1,24]
cGMP-PKG	Signal transduction	65	74	201	12	[1,25]
ErbB	Signal transduction	60	91	309	10	[1,19]
FoxO	Signal transduction	80	78	70	9	[1,25]
Hedgehog	Signal transduction	59	50	41	5	[1,24]
HIF-1	Signal transduction	65	76	54	7	[1,17]
Hippo	Signal transduction	91	85	73	6	[1,24]
JAK-STAT	Signal transduction	85	262	12958	12	[1,25]
MAPK	Signal transduction	119	172	2083	11	[1,21]
mTOR	Signal transduction	76	91	375	14	[1,16]
NF-Kappa B	Signal transduction	137	122	110	6	[1,24]
Notch	Signal transduction	25	30	64	4	[1,20]
Phospholipase D	Signal transduction	56	71	235	11	[2,25]

*Suite du tableau ??*

Nom de la voie	Catégorie KEGG	Nombre de gènes dans la voie	Nombre d'interactions gene-gene	Nombre de sous-voie	Nombre de gènes dans la plus grande sous-voie	Etendue des moments des naissances des gènes
PI3K-Akt	Signal transduction	90	97	668	13	[1,25]
Rap1	Signal transduction	80	99	108	5	[1,23]
Ras	Signal transduction	87	112	232	9	[1,23]
Sphingolipid	Signal transduction	63	72	51	6	[1,23]
TGF-Beta	Signal transduction	73	76	61	7	[1,20]
TNF	Signal transduction	101	54	27	8	[1,24]
VEGF	Signal transduction	28	34	19	10	[1,18]
Wnt	Signal transduction	85	98	402	11	[1,24]

Légende : Ce tableau récapitule les caractéristiques de chacune des 47 voies de signalisation utilisée pour nos deux études. Le nombre de gènes est le nombre de gène unique, les paralogues ne sont pas comptabilisés. L'étendue des moments de naissance des gènes est représentée avec [nœud le plus ancien, nœud le plus récent].

## 1.4 Notions supplémentaires : Question de dosage des gènes

Ce chapitre est une partie secondaire à la thèse, mais nécessaire à la compréhension de l'article 3 en Annexe ??.

L'une des pressions qui peut régir sur le maintien des gènes en forme dupliquée ou leur retour en singleton peut-être une question de dosage entre deux protéines interagissant ensemble. La question de dosage est directement liée à la quantité d'une protéine dans un organisme, et par extension peut être lié au nombre de copies de gènes nécessaires pour produire une protéine fonctionnelle. Dans la majorité des cas un allèle (un exemplaire d'un gène) suffit pour être fonctionnel. Cependant il existe des cas où la pression de la quantité génique est très importante, notons les gènes haplo-insuffisants, mono-alléliques, soumis à empreinte parentale et du chromosome X.

Les gènes haplo-insuffisants sont des gènes pour lesquels une simple copie du gène n'est pas suffisante pour être fonctionnel. Ce qui sous-entend qu'ils ont besoin d'être au nombre de deux pour le devenir. Les gènes haplo-insuffisants, lorsqu'ils ont perdu une de leurs copies peuvent mener à différents phénotypes (individus +/-) (**johson causes 2019**).

Au contraire des haplo-insuffisants, les gènes d'expression mono-alléliques ont a priori besoin d'être en une seule copie pour être fonctionnels. Il y a par exemple les protéines immunoglobulines du système immunitaire qui sont présents en plusieurs formes d'allèle dans le génome, seulement un seul allèle à chaîne légère et à chaîne lourde sont activés (**vettermann allelic 2010**).

Chez les euthériens, les gènes soumis à empreinte parentale sont caractérisés par le fait qu'ils s'expriment uniquement s'ils viennent du père ou de la mère. Quant aux gènes portés par le chromosome X, la majorité sont éteints sur l'un des deux chromosomes, et ce de façon aléatoire, clonale et très tôt chez l'embryon femelle (**balaton exceptional 2018**). Dans notre étude 3, nous nous sommes focalisés sur ces gènes soumis à une pression de dosage pour le clade des téléostéens.

# Article 1 - Moment d'apparition des gènes impliqués dans une voie de signalisation humaine

## 2.1 Contexte de l'étude

La communication cellulaire est un mécanisme important comme expliqué au Chapitre ???. Comme précédemment expliqué, une voie de signalisation s'initie principalement par la fixation d'un ligand sur son récepteur membranaire. C'est donc dans cette logique de poursuivre les travaux menés par Anna Grandchamp lors de son doctorat (2015-2018) qui portait sur le moment d'apparition des ligands et de leurs récepteurs membranaires dans l'arbre de la vie des métazoaires que cette étude a vu le jour. Nous nous sommes intéressés à l'ensemble des voies de signalisation intracellulaire, et nous avons étudié s'il existait un lien entre la position des protéines au sein d'une voie, et le moment d'apparition des gènes correspondants dans l'arbre de la vie. Cette étude est présentée sous forme de schéma récapitulatif en Figure ?? page ?? et est suivie d'un article soumis page ??.

## 2.2 Matériels et méthode

Pour cette étude, comme pour la prochaine, nous avons récupéré un ensemble de 2 298 gènes uniques impliqués dans 47 voies de signalisation humaine dans la base de données KEGG V104. Les voies et leurs caractéristiques sont présentées en ?? Tableau ??.

Afin de déterminer le moment d'apparition, nous avons récupéré les arbres Genomicus Vertebrates V109 et Metazoa V51. Pour chacun de nos gènes, nous avons remonté l'arbre jusqu'à obtenir l'orthologue le plus ancien décrit. Le nœud le plus ancien regroupant l'espèce contenant l'orthologue de l'homme est alors noté comme étant le moment d'apparition du gène en question. Nous récupérons également toutes les interactions protéine-protéine pour regarder l'ordre d'apparition dans la voie, qu'il soit « *backward* » (le membre

impliqué en amont dans la voie (proche du couple ligand-récepteur) est arrivé en premier), « *forward* » (le membre impliqué en aval dans la voie (proche d'un facteur de transcription) est arrivé en premier), ou simultanée (les deux membres sont apparus au même nœud évolutif). Et enfin, pour déterminer une éventuelle corrélation entre la position des protéines dans la voie et le moment de naissance du gène correspondant, nous avons récupéré l'ensemble des portions de voie que nous appelons “sous-voies” d'une voie et attribué un rang à chaque gène en fonction de la position qu'il occupait dans cette sous-voie. Des tests de permutation ont été réalisés pour étudier si les relations *forward*, *backward* et simultanées étaient dues au hasard ou non, ainsi que des tests de corrélation (Pearson) pour vérifier que la « relation position des gènes dans la voie » (près de la membrane ou près de la cible souvent le noyau) et le nœud de naissance du gène dans l'arbre était liés.

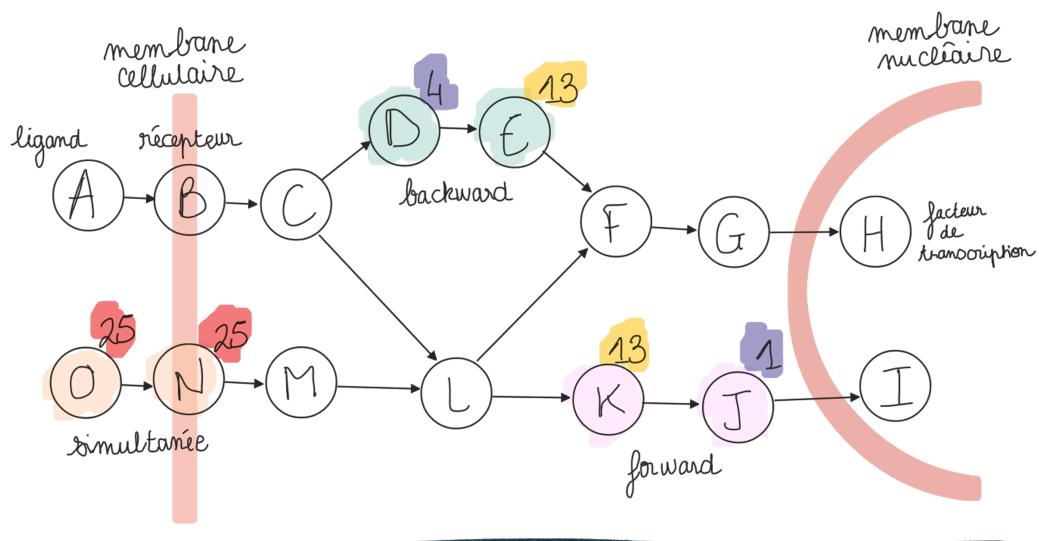
## 2.3 Résultats

Dans un premier temps, nous avons montré qu'il existait 2 principaux nœuds dans l'arbre de la vie à l'origine des gènes impliqués dans les voies de signalisation, à la racine des opistocontes et à la racine des vertébrés. Concernant les interactions gène-gène, les partenaires ont des moments d'apparition le plus souvent asynchrone (81,29%, dont 39,39% de relation *backward* et 41,90% de relation *forward*), contre 18,71% de naissances simultanées pour les deux partenaires d'une interaction. Pour 4 voies, Hedgehog, RIG-I-like receptor, C-type lectine receptor et Thyroid hormone, on retrouve significativement plus d'interactions *forward* qu'aléatoirement. Pour 5 voies, Sphingolipid, AMPK, Notch, Ovarian steroidogenesis et Thyroid hormone, on retrouve significativement plus d'interactions *backward* qu'aléatoirement. Et pour 25 voies, on retrouve plus d'interactions apparues simultanément qu'en prenant des interactions au hasard. Enfin, pour 26 voies, on retrouve une corrélation négative ( $p\text{-value} < 0,015$ ) entre les positions dans la voie et les nœuds de naissance des gènes, ce qui veut dire que ces voies auraient une tendance à s'être construites de l'aval de la voie (facteur de transcription) vers le début de la voie (ligand-récepteur). Pour 10 voies, on retrouve une corrélation positive ( $p\text{-value} < 0,026$ ), ce qui veut dire que ces voies auraient une tendance à s'être construites de l'amont de la voie vers l'aval de la voie.

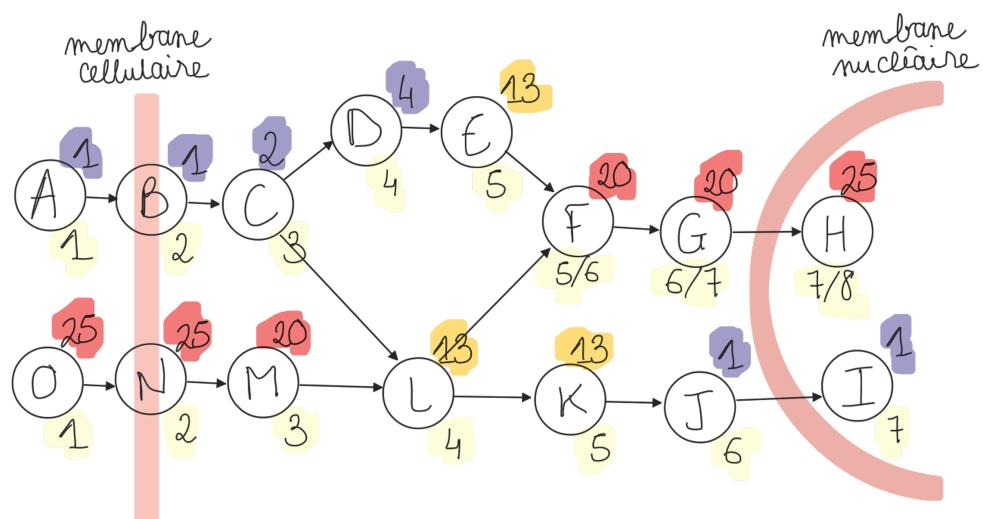
## 2.4 Conclusion

Deux scénarios évolutifs semblent se dégager concernant la “construction” des voies de signalisation humaine au cours de l'évolution et motivées par deux moments clés dans l'arbre de la vie des animaux : à la racine des Opistocontes ainsi qu'à la racine des vertébrés. Pour 26 de nos voies de signalisation, elles se seraient plutôt construites de l'aval de la voie vers l'amont de la voie, et pour 10 autres, ce serait le scénario inverse.

## PARTIE 1



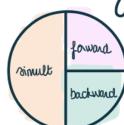
## PARTIE 2



### legende

- protéine
- interaction directionnelle
- 2 position dans la voie
- Quarthocentes 1
- Vertébrés 13
- Mammifères 25
- moment de naissance du gène simplifié

### hypothèses



interactions apparaissent simultanément ++

1

differents sens de construction de la voie

2



FIGURE 2.1 – Schéma récapitulatif de l'étude 1

## 2.5 Article

Genes encoding intracellular signaling proteins in animals  
originated primarily in Opsithokonta and Vertebrata

Floriane Picolo<sup>1</sup>, Jérémie Bardin<sup>2</sup>, Michel Laurin<sup>2</sup>, Benoît Piégu<sup>1</sup>, and Philippe Monge<sup>\*1</sup>

<sup>1</sup> PCR, UMR85, INRAE, CRNS, IFCE, Université de Tours, F-37380 Nouzilly, France

<sup>2</sup> CR2P "Centre de Recherches sur la Paléo-biodiversité et les Paléo-environnements", UMR 7207,  
CNRS/ MNHN, Muséum National d'Histoire Naturelle, Sorbonne Université, Paris

Corresponding author: philippe.monge@inrae.fr

## Abstract

In this work, we evaluate the time of appearance of genes involved in animal signaling pathways with the aim of studying the times of appearance of genes relative to their partners. Signaling pathways are very little described outside humans. We therefore used the human signaling pathways described in kegg (47) as a framework to infer the relationships between genes. Furthermore, we searched for the orthologs of these genes in 315 animal species plus yeast, grouped into 25 clades, in order to determine the time of appearance of each gene.

For this study, 47 human intracellular signaling pathways are used, and the animal life tree to the most common ancestor of vertebrates and yeast, with a species panel of 316 animal species and a tree comprising 25 nested clades to determine the timing of gene onset. Our hypothesis is that there is a link between the time of appearance of a gene and its position in a signaling pathway. We first found that 2 key clades stand out for all genes: Opisthokonta and Vertebrata. In a second step, we observe that for interactions 2 to 2 involved in a signaling pathway, only 16% of genes arose simultaneously, compared to 40% of interacting interactants on a partner arrived before their partner and 43% arrived after their partner. In addition, for 25 pathways, we have a negative correlation between the time of birth of genes and their position in the pathway, that is, the earlier the gene originates in the tree of life, the later it interacts in the signaling pathway.

## Introduction

The co-evolution of genes encoding interacting molecules is a subject of intense study (Andreani & Guerois, 2014; Fraser et al., 2004; Lynch & Hagner, 2015; Rand et al., 2004) because of the intriguing question of the modes of mutation and selection that act on two molecules simultaneously. In particular, the co-evolution of the binding motif has been well investigated (Lewis et al., 2010). These studies of co-evolution focused for example on the fitness (Bloom et al., 2004; Williams et al., 2001), on the conservation of the interaction (Kachroo et al., 2015; Lovell & Robertson, 2010; Wuchty et al., 2003), or on the evolution of the residues at the interface of the molecules (Echave & Wilke, 2016; Jack et al., 2016; Mintseris & Weng, 2005). While these studies on the coevolution of binding partners often require the integration of different disciplines (chemistry, evolution, biology), the establishment of the interaction from a phylogenetic point of view is less studied. Little is known for example about the origin and evolution of the different partners prior to their first interaction. Does the emergence of one partner favor the emergence of the second partner?

In the case of interacting molecules, the appearance of genes coding for molecules included in a complex is more intricate (Kauffman, 1993). For two molecules that will eventually interact, the appearance of one may be dependent on the appearance and conservation of the other. This may be the case, for example, when the presence of the first molecule is not advantageous if its partner has not yet appeared (a sort of exaptation).

The existence of interacting proteins without partners (“orphan partners”) has been frequently described (Howard et al., 2001), even though it is sometimes difficult to assess whether an interacting protein is a true orphan or its ligand is just unknown (Benoit et al., 2006). The relative order of appearance of genes encoding protein partners is thus an open question. Furthermore, several types of interactions can be observed in living organisms, with different numbers of interacting partners (Albelda & Buck, 1990; Koshland, 1958; Maslov & Sneppen, 2002; Sullivan & Holyoak, 2008), varying affinities (Kent et al., 1980), or different duration for the interactions (Nooren & Thornton, 2003), making the problem more complex.

Concerning the pairs of ligand and its receptor, Thornton (2001) has shown that the first steroid receptor of the family, present in lamprey and supposed to be present in the last common ancestor of vertebrates, was an oestrogen receptor, and that several duplications led to other steroid receptors, specialized in other functions with other ligands. However, more recent investigations suggested that the ancestral ligand for the ancestral steroid receptor was a molecule with a structure distinct from modern oestrogen, an aromatized steroid with a side-chain, called paraestrol (Markov et al., 2017). Concerning membrane receptors, in a previous work, we found that 41% of the receptors and their respective first ligands appeared on the same branch of the evolutionary tree, representing 2.5-fold more than expected by chance, thus suggesting an evolutionary dynamic of interdependence and conservation between these

partners (Grandchamp & Monget, 2018). In contrast, 21% of the receptors appeared after their ligand, i.e., three-fold less often than expected by chance, and 38% of the receptors appeared before their first ligand, as much as expected by chance. These results suggest that a selective pressure is exerted on ligands and receptors once they appear, that would remove molecules whose partner does not appear quickly.

In the present work, we investigated the node of appearance of genes encoding proteins involved in signaling pathways, downstream of membrane receptors in the animal tree of life (Figure 1). We also studied the relationship between the position of the protein in the signaling pathway relatively to his partners, and the node of appearance in the tree of life. In other words, does the order of appearance of genes in the tree of life reflect that of their coded protein in the signaling pathway (upstream or downstream of the pathway), with a direct or reverse order (Figure 2)?

## Results

### The nodes of origin of genes encoding proteins of signaling pathways

First, we looked at the distribution of birth times for each of our genes encoding proteins involved in an animal signaling pathway (2,298 unique genes involved in 47 human intracellular signaling pathways available on the KEGG database V104.0, see *M&M*). Two parts of the tree of life are overrepresented in gene origins (Figure 3): Yeast and Porifera at the base of the tree, which can be more broadly encompassed in *Opisthokonta* (13000 MYA), and the *Tunicata*, *Petromyzontidae*, *Mixini*, *Chondrichthyes* and *Actinopterygii*, which correspond to the first diversification of Vertebrates (758 MYA) (Kumar et al., 2022).

Moreover, we looked at each pair of proteins interacting with each other to study whether both appeared at the same time/node in the tree of life, or whether one arrived before the other, and if so, the length of the delay. Of the 3,000 interactions studied, 16.3% of interactions are “simultaneous”, i.e., both partners in an interaction appeared at the same clade, and 83.7% are asynchronous, of which 43.2% are “forward”, i.e., partner 2 (downstream in the pathway) arrived first, 40.5% are “backward”, i.e., partner 1 (upstream in the pathway) arrived first (Figure 4). Some pathways have more backward relationships than others, such as the MAPK and IL-17 pathways in comparison with the "Ovarian steroidogenesis" pathway, in proportion to their respective numbers of interactions. It is the opposite for other pathways such as the PPAR and cAMP pathways.

Secondly, simulation procedures allowed us to compare the relative quantities of interaction of each type (forward, backward or simultaneous) to distributions obtained by randomly mixing the relationships between the elements of the pathways. For 4 pathways, Hedgehog, RIG-I-like receptor, C-type lectin receptor and Thyroid hormone (Suppl. Data 1), there are significantly more forward interactions ( $>56.20\%$ ) than if these interactions were taken at random ( $p < 0.05$ ). For 5 other pathways (Sphingolipid, AMPK, Notch, Ovarian Steroidogenesis and Thyroid hormone), backward interactions were found more often ( $> 50.3\%$ ) than randomly ( $p < 0.05$ ), and in particular for the Ovarian Steroidogenesis pathway for which in 91.5% of the tests we found more backward interactions than random ones. Finally, concerning the Hippo pathway, in 84.2% of the tests, we found significantly more interactions appearing simultaneously than by taking interactions at random, and this is the case for 24 other pathways ( $>53.6\%$ ;  $p < 0.05$ ). For 13 pathways, we did not find a significant difference between the chance of establishing one of the three types of interaction (backward, forward or simultaneous) and establishing it at random.

We then looked at the delta between the node of birth of the first interactant and that of the second interactant for the 47 pathways studied (Figure 5). For 12 pathways (at the top of Figure 6), the delta median is greater than 0 (between 0.5 and 3), for 6 pathways the median is less than 0 (between -2 and -1), and for 29 pathways the median is 0. Moreover, certain pathways have specific profiles. Firstly, the

PPAR pathway exhibits a very restricted boxplot with very minimal variance (min = -1, Q1 = -0.50, median = 0, Q3 = 0.5 and max = 1). And some pathways exhibit median particularly eccentric to 0 such as the IL-17 pathway (min = -12, Q1 = -8, median = -2, Q3 = 0.5 and max = 12), and the “ovarian steroidogenesis” pathway (min = -6, Q1 = 0.5, median = 3, Q3 = 5.5 and max = 14) (Q1 and Q3 being 1 and 3rd quantiles of a dataset).

For each pathway, we also studied the delta depending on the node of appearance in the tree of life of each partner. For the Hippo pathway, for example, the distribution curve resembles a normal distribution, the interactions having appeared in the same clade (delta = 0) (13.6%) (to the Porifera clade, clade 2 (45.5% of total interaction with delta = 0, (Figure 6 deciphering the line Hippo of the figure 5)). For the Ovarian steroidogenesis and cAMP pathways, the deltas are greater than 0 (median = 3 and 1, respectively), the last elements of the pathway have arrived before terminal elements (Suppl Data 2). The other pathways are represented in Suppl. data 2.

### **Relationship between the node of birth of a partner involved in each pathway, and the upstream or downstream position in the pathway**

For 25 of the 47 pathways studied here, there is a negative correlation (p-value range from 3.73E-104 to 0.02) between the node of birth of a gene and the position of the corresponding protein in the pathway, which means that upstream proteins in a signaling pathway tend to appear late in the tree of life, for instance in vertebrates to eutherians (Hypothesis 2 of Figure 2 B, Table 1 and Figure 8). For three of them, this correlation is particularly strong: Adipocytokine ( $r = -0.58$  and  $p = 2.44E-24$ ), Fc epsilon RI ( $r = -0.53$  and  $p = 2.02E-16$ ), and VEGF ( $r = -0.53$  and  $p = 3.15E-8$ ). For 10 pathways, the correlation between the node of appearance and the position in the pathway is positive (p-value range from 2.79E-280 to 0.03), which means that the more upstream the protein is involved in the pathway, the older the corresponding gene is (node of appearance in Opisthokonta and Metazoa for example). All the correlations are presented in Table 1.

The moment of appearance of each member of a pathway on the KEGG pathways shows interesting patterns (Figure 8 and Suppl. Data 4). In the Hippo pathway (Figure 8 A), there are a lot of ancestral members, distributed throughout the pathway, upstream and downstream (hypothesis 3, Figure 2 C). In the Hippo and the Wnt pathways (Figure 8 A), a lot of partners appear between clades 1 and 2 (Opisthokonta or Metazoa): Mob, TEAD (clade 1), Mer, SAV1, Mst1/2 (clade 2) and YAP/TAZ (clade 3). And the downstream members of these pathways (after the nucleus) arose in vertebrates (clade 12). In the PPAR pathway (Figure 8 B), there are 3 central elements, represented by PPAR $\alpha$  (clade 12),  $\delta$  (clade 12) and  $\gamma$  (clade 14), dimerized with RXR ( $\alpha$  (clade 13),  $\delta$  (clade 15) and  $\gamma$  (clade 17)). The PPAR pathway corresponds to hypothesis 2 (Figure 2 B), upstream proteins in the signaling pathway tending to appear late in the tree of life. We also observe examples of our hypothesis 3 (Figure 2 C) with the Notch pathway for example (Figure 8 C), there is not necessarily an order that emerges, and we do not

find any correlation between the positions in the pathway and the clade of birth of corresponding gene. We also have examples of our hypothesis 3 (Figure 2 C) with the Notch pathway for example (Figure 8 C), there is not necessarily an order that emerges, and we do not find any correlation between the positions in pathways and times of gene appearance of proteins in pathways and times of gene appearance ( $r = 0.02$ ,  $pvalue = 0.77$ ). We also observe that no pathway corresponds to hypothesis 1 (Figure 2 A), upstream proteins in a signaling pathway never appearing late in the tree of life (Suppl.data 4).

## Discussion

In our previous work (Grandchamp & Monget, 2018), we have shown that the genes encoding the pairs of ligand/membrane receptor mainly appeared at the root of metazoa, vertebrates and teleosts. In the present work, we show that the genes encoding proteins of signaling pathways mainly appeared in Yeast and Porifera clade in the one hand, and in the node of Vertebrates on the other hand. Moreover, as expected, among all the pathways, at least 218 proteins (from 1 to 33 by pathway) appeared in yeast, showing, as expected, that numerous genes encoding proteins of signaling pathways appeared long before membrane receptors and their ligands. Our previous study on membrane receptors and their ligands was done on a tree with only 10 nodes, and it could be refined by using the 25 nodes as in the present paper.

In the signaling pathways studied, there are therefore numerous proteins that interact with their partner(s) in humans, appearance of both being desynchronized in the tree of life. This raises the question of the functioning of these pathways without the full set of his members. For example, p53, which appeared in Metazoa, is inhibited by several factors whose genes appeared later, such as Mdm2 and Mdm4 (Olfactora). Without these inhibitors, individuals would die *in utero*, as shown in the mouse (Jones et al., 1995; Montes de Oca Luna et al., 1995). This suggests there was another p53 inhibitor before Olfactora appeared. In addition, human p53 targets also emerged later in evolution, such as IGFBP3 (Vertebrates) and pro-apoptotic factor bax (Olfactora). It is these and other factors that explain the tumor suppressor role of p53 in humans (Lehmann-Che et al., 2007) but also in *Drosophila* (Zhou, 2019). We can therefore hypothesize a progressive refinement of the mechanisms of action and inhibition of p53 during evolution.

Another example is the interaction between PIN1 and IRF3 in the RIG-I-like receptor pathway, which have a delta of 13, PIN1 gene having appeared at clade 1 (Yeast) and IRF3 at clade 14 (Chondrichthyes). Pin1 also activates p53 which appeared on the branch leading to clade 2 (node of Metazoa) (Berger et al., 2005). Another example of desynchronization concerns the Sphingolipid pathway, in which CTSD interacts only with BID, and BID interacts with CTSD and BAX in humans. CTSD appeared in Metazoa, BAX in Olfactora, and BID gene in Prototheria. Moreover, although yeast genome does not contain genes encoding Bcl-2 proteins, the heterologous expression of mammalian Bax in yeast induces a suppressible lethal phenotype that is associated with characteristics of metazoan apoptosis, strongly suggesting that its targets are already present in yeast (Khoury & Greenwood, 2008; Zha et al., 1996).

We found that for 33 (4+5+24) pathways, there are significantly more forward, backward or simulated interactions that have been established than if these interactions were taken randomly. This suggests that for 14 pathways, the directions of these interactions were established by chance. In other words, for these interactions, they seem to have been established by chance, and probably because it worked in the cell, these interactions were conserved during evolution.

In a previous study, we described the cases where the genes coding for membrane receptors appeared before their ligand (Grandchamp & Monget, 2020). We had studied more precisely the 30 cases for which the 3D structures were known in the PDB database, to formulate hypotheses on the plausible scenarios of evolution of the amino acids involved in the binding pocket of the receptor, until the ligand appeared. In the present work, a similar study would be very interesting for the partners which appeared before the proteins with which they interact. Such a study could be performed for ras/sos complexes for example, grb2/sos, akt/gsk3 or akt/mTor, for which the 3D structure is available.

Some signaling pathways are relatively old (i.e., contain more members that appeared in yeast or metazoans than in more recent clades), such as the MAPK pathway, other pathway being more recent, such as the pathways involved in immunity. Of note, genes involved in immune pathways evolve very quickly compared to the rest of the genome (Cooper & Alder, 2006).

Almost all signaling pathways have target factors (e.g., factors entering the nucleus) that appeared very early during evolution (dark blue factors to the right in the colored KEGG diagrams, Figure 8 and Suppl data 3). However, for certain pathways, the factors that appeared earliest (yeast or metazoa) are close to the targets, and those further upstream in the pathway appeared later (PPAR, cAMP), whereas for other pathways (IL-17, T cell receptor), certain factors close to the targets appeared later than the factors acting upstream of the pathway. This shows that pathways were not formed/refined in the same way, nor according to the same timing during evolution.

This study could have been completed by an analysis based on known pathways in yeast to validate or invalidate certain interactions that are considered to exist if the two genes are present. However, only one pathway from our list, MAKP, is described in KEGG for the *Saccharomyces cerevisiae* species, this pathway being widely documented (Chavel et al., 2014; Saito, 2010; Zou et al., 2008), including in plants (Meng & Zhang, 2013). In the literature, other pathways are also documented in yeast, such as cAMP (Portela & Rossi, 2020; Tamaki, 2007), mTOR (Powers et al., 2004), Ras (Tisi et al., 2014) or Sphingolipid (Montefusco et al., 2014). Thus, using the KEGG database, it is not possible to establish an exhaustive evolutionary bridge of signaling pathways between yeast and humans. Moreover, some species lack certain pathways, such as yeast, which lacks the NF-kappa B pathway (Ho et al., 2017; Saleski et al., 2017), and yet some elements of the pathway are present in *Saccharomyces cerevisiae* such as caseins kinase 2 (CSNK2A1/2/3 and CSNK2B). In yeast, these caseins are known to be essential for mitophagy (Kanki et al., 2013).

Another point of discussion concerns the methodology of the dating of the appearance of a gene. For this we are limited by the different versions of Genomicus and Genomicus Metazoa. In particular, the “intersection species” between these two databases are *Drosophila melanogaster* and *Caenorhabditis elegans*. Indeed, if for a given gene the oldest ortholog found is not one of these two species, or if this gene is lost in both species but present in more ancient taxa, our methodology does not allow us to use

adequately the Genomicus Metazoa trees. Genomicus trees are modified trees of Ensembl, and therefore the limit comes from its origin Ensembl. Furthermore, concerning Ensembl, depending on the versions, there may be mega-trees with all the paralogs of a gene in the same tree, or in sub-trees, with one paralog per sub-tree. This change from mega-trees into several sub-trees in Ensembl V94 and further (Emily, 2018) –which is not automatic, it depends on the size of the paralog family–, makes the attribution of appearance times more complex.

Among the 25 clades that we have selected, not all have been studied to the same extent. As shown in Suppl Data 4, we can see that Yeast (clade 1) only includes one species (*Saccharomyces cerevisiae*) while for Aves (clade 19), we included 13 species. However, our data show that appearance of the relevant genes is concentrated on 2 nodes (those subtending Opisthokonta, and Vertebrates) and the genes that appeared on these branches represent 75.7% of all the genes involved in the KEGG pathways.

It must also be considered that the signaling pathways noted on KEGG (but more broadly on databases) are human representations, and simplifications have been made to simplify reading and understanding. However, as seen in Figure 8 A, some proteins are involved in multiple pathways, such as the Shc → Grb2 → SOS → Ras → Raf1 → MEK → ERK subpathway which is involved (partially or entirely) in 22 pathways (ERBB, Estrogen, GnRH, Insulin, Prolactin, Relaxin, B cell receptor, Chemokine, F epsilon RI, T cell receptor, Neurotrophin, cAMP, FoxO, JAK-STAT, MAKP, mTOR, Phospholipase D, PI3K-Akt, Rap1, Ras, Sphingolipid, VEGF). The RTK/RAS/ERK component is indeed known to be common to *Drosophila*, nematode and humans (Ashton-Beaucage & Therrien, 2010).

## Conclusion

This study highlights two key phases of *Opisthokonta* evolution concerning the genes encoding proteins involved in signaling pathways in animals, near the base of *Opisthokonta* and of Vertebrata. We also observe a correlation (positive or negative depending of the pathway) between the position of the proteins in each pathway and the node of birth of their corresponding gene in the tree of life.

## Material and methods

### Implementation of the database

#### *Signaling pathways by KEGG*

We followed the methodology described in our precedent paper (Picolo et al., in press).

We use the KEGG pathways because they are annotated in humans and therefore the genes are annotated in humans, and we use parsimony and the trees to locate the branch on which each gene originated. We retrieved a list of 2,298 unique genes encoding proteins involved in the 47 human intracellular signaling pathways available on the KEGG V104.0 database (<https://www.genome.jp/kegg>) (Bader et al., 2006) using the keywords “signaling pathway” and “human” (Table II). KEGG is one of the most referenced and used databases listing signaling pathways.

Each signaling pathway was retrieved in .xml (Extensible Markup Language) format from KEGG's PATHWAY Database tool. This file format is well supported by R libraries such as XML (Lang & Kalibera, 2023) and igraph (Csárdi et al., 2023; Csárdi & Nepusz, 2006).

The gene products of KEGG pathways are inscribed in rectangular blocks that we call "labels", for the XML file, this is the label "name", and for simplicity for readers, these labels can cover several paralogs (Figure 9). Some paralogous genes appear under different labels, as with the PPAR pathway for which the different proteins PPAR $\alpha$ , PPAR $\gamma$  and PPAR $\delta$  have similar interactions (with FABP3, Thiolase B, aP2 and UCP-1) and specific interactions (PPAR $\alpha$  → HMGCS2; PPAR $\gamma$  → GyK for example). Each pathway is constructed in such a way that there can be one or more inputs (e.g., a ligand) and one or more outputs (e.g., transcription factor). Each sub-pathway is analyzed, which means that the same gene can be involved in several sub-pathways (e.g.: A → B → C and A → D → E).

#### *Tree of life from Genomicus*

To determine the time of appearance, two phylogenetic trees from Genomicus (Nguyen et al., 2018) were used: the vertebrate tree and the metazoan tree. The vertebrate tree is the V109 tree (<https://www.genomicus.bio.ens.psl.eu/genomicus-109.01>), comprising 199 animal species and covering a range of species belonging to the vertebrate clade, as well as *Drosophila melanogaster* and *Caenorhabditis elegans* and to this is added a yeast : *Saccharomyces Cerevisiae*. The metazoan tree is that of V51 (<https://www.genomicus.bio.ens.psl.eu/genomicus-metazoa-51.01>), comprises 116 animal species and covers all main taxa belonging to the non-vertebrate metazoan clade. *Drosophila melanogaster* and *Caenorhabditis elegans* are the two “intersection species” between these two databases. (Figure 1, Suppl. Data 4).

From these Genomicus trees, we determined 25 clades on a similar and “enriched” model of our previous work (Grandchamp & Monget, 2018): 1 : Yeast (~ 1110 my), 2 : Porifera (~ 725 my), 3 : Placozoa (~ 550 my), 4 : Ctenophora (~ 109 my), 5 : Cnidaria (~ 588 my), 6 : Xenacoelomorpha (~ 550 my), 7 :

Spiralia (~ 610 my), 8 : Ecdysozoa (~ 653 my), 9: Echinodermata (~ 596 my), 10 : Cephalochordata (520 my), 11: Tunicata (~ 446 my), 12 : Petromyzontidae (~ 416 my), 13: Myxini (~ 360 my), 14: Chondrichtyes (~ 413 my), 15: Actinopterygii (~396 my), 16: Coelacanthidae (~350 my), 17: Amphibia (~ 319 my), 18: Testudines (~ 194 my), 19: Aves (~ 109 my), 20: Crocodylia (~ 87 my), 21: Squamata (~ 189 my), 22: Sphenodontia (~ 200 my), 23: Prototheria (~ 166 my), 24: Eutheria (~ 98 my), 25: Metatheria (~ 66 my). To provide consistent ages, and given that several of the oldest nodes are poorly constrained by the fossil record, we have used molecular time estimates throughout (de Vienne, 2016; Kumar et al., 2022; Nguyen et al., 2021). These clades are our references for the present study. The gene is considered to appear on the branch that leads to the clade for which all included taxa possess the same gene.

#### Dating the appearance of genes

We consider the node of appearance by the presence of an ortholog of our gene within clades (e.g.: gene A has an ortholog shared by humans and *Saccharomyces cerevisiae*, so its appearance is assumed to predate Opisthokonta) (<https://github.com/florianepicolo/birth-gene>).

Some cases require a more complex analysis. If under a label in the KEGG pathways, several paralogues are listed, then the latest possible date of origin of this gene is the oldest node at which paralogs occur. If several paralogs are present in a pathway but under different labels, we considered them independently from other paralogs. In the case of a protein complex, we considered the complex to be a group, so we considered the node of birth of the complex the deepest node for which all the genes of the complex are present, considering that the complex cannot be functional unless all the proteins are present.

All elements in a pathway are assigned a rank in the pathway, and multiple ranks are possible for the elements that are present in several sub-pathways. We have not considered the “components” of the pathways present on KEGG in the allocation of the ranks if they are not proteins, and therefore do not involve genes, like elements like Ca<sup>2+</sup> or lactate (example: A<sub>1</sub> → B<sub>2</sub> → Ca<sup>2+</sup> → C<sub>3</sub>, with ranks in index).

Furthermore, in an interaction A → B (A is closer to the receptor than B), the interaction is said to be “forward” if gene A was born after B, “backward” if gene A was born before B, and “simultaneous” if A and B were born on the same branch.

In the present study, ten pathways are involved in immunity (Table II), their genes appearing later (from the vertebrates) in the tree of life than the average (chi test p = 3.50E-08)

#### Analyses and statistics

To determine whether there is more forward, backward, or simultaneous appearance interaction compared to random interactions, we performed permutation test (1000 permutations). We performed

correlation test (Pearson) to determine a possible relation between the branch of appearance of a gene and the position (upstream or downstream) of the corresponding protein in the pathway.

The distribution of relationships also needs to be considered. Indeed, the relationships of the same types (i.e. backward, forward, and simultaneous) can be spread in the pathways or clustered. To answer whether these clusters are bigger or smaller compared to randomness, we adopted a simulation approach described hereafter and for which a visual flowchart is available in sup. Mat X. For each pathway, we extracted the components of the five different types: forward (target younger than origin), forward and simultaneous (target's age equal or lower than origin), simultaneous (equal age), backward (target older than origin), backward and simultaneous (target's age equal or higher than origin). For each five types of relationships, we extracted the distribution of these clusters sizes. We then simulated 1000 alternative pathways under the null hypothesis ( $H_0$ ) that the relationships between genes have no relationship with their age. We did that simply by permuting relationships between genes without altering ages. For each simulated pathway under  $H_0$ , we obtained five corresponding distributions of clusters sizes. To verify if these latter have bigger or smaller clusters sizes, we did a permutation test based on means. We finally considered how frequent random pathways have smaller or bigger clusters sizes, how much is the difference between these distributions (raw vs simulated under  $H_0$ ) and how frequent it is significant.

**Table I - Correlation between time of birth and rank in the lane**

Pathway	Number of subpath	Max subpath	Birth min, max	Correlation (r)	pvalue
Adipocytokine	54	8	[1,17]	<b>-0.58</b>	<i>2.44E-24*</i>
Fc epsilon RI	42	10	[1,24]	<b>-0.53</b>	<i>2.02E-16*</i>
VEGF	19	10	[1,18]	<b>-0.53</b>	<i>3.15E-08*</i>
PPAR	61	2	[1,23]	-0.48	<i>4.41E-08*</i>
Insulin	255	11	[1,25]	-0.42	<i>1.28E-78*</i>
Rap1	108	5	[1,23]	-0.39	<i>2.01E-16*</i>
B cell receptor	88	13	[1,22]	-0.38	<i>2.76E-19*</i>
C-type lectin receptor	107	8	[1,25]	-0.36	<i>7.24E-13*</i>
Thyroid hormone	102	7	[1,21]	-0.34	<i>2.11E-11*</i>
Sphingolipid	51	6	[1,23]	-0.32	<i>1.26E-06*</i>

Estrogen	29	16	[2,24]	-0.31	<i>1.14E-04*</i>
Apelin	26	11	[1,21]	-0.31	<i>4.36E-06*</i>
TNF	27	8	[1,24]	-0.29	<i>4.15E-04*</i>
ErbB	309	10	[1,19]	-0.26	<i>6.67E-31*</i>
cGMP-PKG	201	12	[1,25]	-0.23	<i>2.76E-16*</i>
Glucagon	40	9	[1,25]	-0.22	<i>2.92E-03*</i>
Oxytocin	51	9	[1,21]	-0.22	<i>1.88E-04*</i>
mTOR	375	14	[1,16]	-0.21	<i>9.75E-31*</i>
HIF-1	54	7	[1,17]	-0.20	<i>1.38E-02*</i>
GnRH	16	16	[1,21]	-0.19	<i>2.00E-02*</i>
MAPK	2083	11	[1,21]	-0.18	<i>3.73E-104*</i>
Phospholipase D	235	11	[2,25]	-0.18	<i>3.61E-14*</i>
TGF-beta	61	7	[1,20]	-0.15	<i>2.05E-02*</i>
Neurotrophin	952	16	[1,25]	-0.13	<i>8.28E-35*</i>
Chemokine	183	11	[1,21]	-0.10	<i>1.52E-04*</i>
Ras	232	9	[1,23]	0.06	<i>3.46E-02*</i>
NOD-like receptor	420	9	[1,25]	0.06	<i>7.44E-04*</i>
cAMP	1966	11	[1,24]	0.07	<i>6.55E-14*</i>
IL-17	709	9	[1,25]	0.07	<i>2.12E-05*</i>
T cell receptor	376	9	[1,25]	0.08	<i>2.33E-04*</i>
JAK-STAT	12958	12	[1,25]	0.11	<i>2.79E-280*</i>
NF-kappa B	110	6	[1,24]	0.17	<i>1.55E-04*</i>
Wnt	402	11	[1,24]	0.22	<i>6.55E-29*</i>
Calcium	77	8	[1,21]	0.36	<i>1.16E-09*</i>
RIG-I-like receptor	374	8	[1,19]	0.49	<i>3.01E-142*</i>

Ovarian steroidogenesis	15	6	[1,25]	-0.22	<i>6.13E-02*</i>
FoxO	70	9	[1,25]	-0.10	<i>8.54E-02*</i>
PI3K-Akt	668	13	[1,25]	-0.02	1.62E-01
p53	254	7	[1,25]	0.04	1.76E-01
AMPK	184	10	[1,25]	0.04	2.70E-01
Toll-like receptor	371	10	[1,25]	0.02	2.97E-01
AGE-RAGE	358	6	[1,24]	-0.02	4.53E-01
Hedgehog	41	5	[1,24]	-0.05	6.23E-01
Relaxin	140	13	[1,24]	0.01	6.28E-01
Notch	64	4	[1,20]	0.02	7.66E-01
Prolactin	63	14	[1,25]	-0.01	7.96E-01
Hippo	73	6	[1,24]	0.01	9.03E-01

The correlation index ( $r$ ) quantifies the strength of the relationship, it is said to be strong when it is between [-1, -0.5] and [0.5, 1] (in bold on Table 1), and it is said to be low between [-0.5, 0[ and ]0, 0.5], and without correlation to  $r = 0$ . Significant values (\*,  $p < 0.05$ ) are written in italics.

**Table II - List of signaling pathways and their characteristics**

<i>Signaling pathway</i>	<i>KEGG categories</i>	<i>Number of genes</i>	<i>Number of interactions</i>
p53	Cell growth and death	64	70
AGE-RAGE	Endocrine and metabolic disease	62	93
Adipocytokine	Endocrine system	36	48
Estrogen	Endocrine system	62	69
Glucagon	Endocrine system	50	55
GnRH	Endocrine system	41	39
Insulin	Endocrine system	62	77
Ovarian steroidogenesis	Endocrine system	45	25
Oxytocin	Endocrine system	58	73
PPAR	Endocrine system	51	63
Prolactin	Endocrine system	54	55
Relaxin	Endocrine system	81	103
Thyroid hormone	Endocrine system	78	85
B cell receptor	Immune system	47	57
C-type lectin receptor	Immune system	154	185
Chemokine	Immune system	58	82
FC epsilon RI	Immune system	42	47
IL-17	Immune system	91	152
NOD-like receptor	Immune system	168	164
RIG-I-like receptor	Immune system	53	73
T cell receptor	Immune system	66	99
Toll-like receptor	Immune system	79	109
Neurotrophin	Nervous system	77	117
AMPK	Signal transduction	69	67
Apelin	Signal transduction	62	78
Calcium	Signal transduction	52	67
cAMP	Signal transduction	88	120
cGMP-PKG	Signal transduction	65	74
ErbB	Signal transduction	60	91
FoxO	Signal transduction	80	78
Hedgehog	Signal transduction	59	50
HIF-1	Signal transduction	65	76
Hippo	Signal transduction	91	85
JAK-STAT	Signal transduction	85	262
MAPK	Signal transduction	119	172
mTOR	Signal transduction	76	91

NF-Kappa B	Signal transduction	137	122
Notch	Signal transduction	25	30
Phospholipase D	Signal transduction	56	71
PI3K-Akt	Signal transduction	90	97
Rap1	Signal transduction	80	99
Ras	Signal transduction	87	112
Sphingolipid	Signal transduction	63	72
TGF-Beta	Signal transduction	73	76
TNF	Signal transduction	101	54
VEGF	Signal transduction	28	34
Wnt	Signal transduction	85	98

### Legends figures

#### Figure 1 – Simplified animal tree of life and clades of study

Each rectangle represents a node of speciation during evolution. Each branch represents a terminal clade of the tree with its associated number in the colored circles. The length of the branches is not representative of the actual evolutionary divergence. In bold are the clades available on Genomicus Vertebrates, and in italics are the clades available on Genomicus Metazoa.

#### Figure 2 - Schematic representation of three possible scenarios linking the order of appearance of a gene and the position (rank) of the corresponding protein in the pathway

A–C represent the three possible scenarios. The colors represent branches of birth for a gene (from 1 to 25). We arbitrarily chose ten proteins/positions per pathway for the illustration. In scenario A, the order of proteins in the pathway matches the order of appearance of genes in the tree. Scenario B is characterized by the opposite relationship (reverse order between protein position in the pathway and of the genes coding these proteins). In the scenario C, there is no link between the position of the protein in the pathway and the order of appearance of the corresponding gene.

#### Figure 3 - Distribution of birth of genes encoding proteins involved in the 47 signaling pathways on the branches subtending 25 nodes

Distribution of nodes of birth for each of the genes encoding proteins involved in a signaling pathway. In the text, these are sometimes referred to by number (1 being Opisthokonta, not shown here), which matches the order shown here (i.e. Methateria is node 25).

#### Figure 4 – Backward, forward or simultaneous interaction for each signaling pathway

Each horizontal column represents a signaling pathway (described on the left). On each column, for a A → B interaction, the orange color corresponds to simultaneous interaction (A was born at the same node

as B), the light green color corresponds to forward interaction (A was born after B), the green color corresponds to backward interaction (A was born before B).

**Figure 5 - Distribution of differences (delta) in the node of birth of genes encoding two partners for each pathway**

Boxplot distribution for each pathway of the difference (delta) between the positions (as rank order, disregarding absolute age) of the node of birth of a gene encoding one protein and of the node of birth encoding one of its partner(s). The pathways are centered on their median (dark vertical line in the center of the box), representing the middle half of the data between the first quartile (Q1) and the third quartile (Q3). Horizontal lines extend from the box to show data dispersion of the data, and isolated points indicates outliers.

**Figure 6 - Distribution of deltas of node of birth of genes encoding proteins involved in the Hippo pathway, according to the node of birth of each gene**

Deltas are calculated via number of clade of birth for the A gene - clade of birth for the B gene for the interaction A → B. For example, if gene A was born on the branch below the blue clade (clade 1), and its delta with B is -10, then clade of birth of B is 11. Moreover, in this case, it is a backward relationship, because A was born before B. The distributions for all the other pathways are shown in (Suppl. Data 1).

**Figure 7 – Node of birth of genes encoding proteins involved in each pathway according to the position (rank) of the protein in the pathway**

Each horizontal column corresponds to one pathway. On the left, the upstream proteins of the pathway (close to the cell membrane), on the right, the downstream proteins of the pathway (close to the nucleus). Depending on signaling pathways, there are between two to sixteen ranks on the abscissa. Each rectangle represents for a given rank the distribution of nodes of birth for each of the proteins occupying the rank. For example, for the GnRH pathway at position/rank 1, all the proteins have appeared at clade 14 (*Euteleostomi*). The pathways are sorted by correlation index in Table I. The distribution of positions/ranks and births of the different proteins for each pathway is accessible in (Suppl. Data 2).

**Figure 8 – Examples of colored KEGG pathways depending on the node of birth of each protein**

Each color represents a clade. The white rectangles correspond to the genes for which we have not been able to determine the node of birth due to lack of information about the gene. The KEGG legend is available here: [https://www.genome.jp/kegg/document/help\\_pathway.html](https://www.genome.jp/kegg/document/help_pathway.html). (A) Hippo signaling pathway; (B) PPAR signaling pathway; (C) Notch signaling pathway. All colored pathways are available in (Suppl. Data 3).

**Figure 9 - Different labels in the KEGG database (example of the Notch pathway)**

A: Graphic representation of the Notch pathway on the KEGG web page. Each rectangle represents one or more proteins of the pathway; the arrows represent an activation, whereas lines that do not end with an arrow but with another short perpendicular represent an inhibition, expression...). B: Extract from the Notch pathway .xml file. For this example, Fringe, which corresponds to the FGN gene corresponds to “id 33” in the .xml file, and we note that there are 3 hsa identifiers, i.e., 3 paralogs: hsa:3955 (LFNG), hsa:4242 (MFNG) and hsa:5986 (RFNG). Next, PSE2, which corresponds to the PSENEN gene, corresponds to “id 26” that has no other paralog involved in this interaction.

### Supplementary data

All additional data are sorted in the following order: p53, AGE-RAGE, Adipocytokine, Estrogen, Glucagon, GnRH, Insulin, Ovarian steroidogenesis, Oxytocin, PPAR, Prolactin, Relaxin, Thyroid hormone, B cell receptor, C-type lectin receptor, Chemokine, FC epsilon RI, IL-17, NOD-like receptor, RIG-I-like receptor, T cell receptor, Toll-like receptor, Neurotrophin, AMPK, Apelin, Calcium, cAMP, cGMP-PKG, ErbB, FoxO, Hedgehog, HIF-1, Hippo, JAK-STAT, MAPK, mTOR, NF-Kappa B, Notch, Phospholipase D, PI3K-Akt, Rap1, Ras, Sphingolipid, TGF-Beta, TNF, VEGF, Wnt.

#### **Suppl. Data 1 - Distribution of deltas of node of birth of genes encoding proteins involved in all the pathways, according to the node of birth of each gene**

Deltas are calculated via clade of birth rank for the A gene - clade of birth rank for the B gene for the interaction A → B. For example, if gene A was born at the blue clade (clade 1), and the clade of birth of B is 11, the A → B delta is -10. Moreover, in this case, it is a backward relationship, because A was born before B.

#### **Suppl. Data 2 - Distribution of genes by position/rank in the pathway and node of birth**

Each graph represents one of the 47 pathways. Abscissa: the different proteins involved in the pathway; ordinate: position/rank of the protein within the pathway. Each protein is colored depending on the node of birth of its corresponding gene. Proteins are represented by a dot, are characterized by their position(s) they occupy within the pathway. The size of the dots is proportional to the number of times they are in these positions.

#### **Suppl. Data 3 - Colored KEGG pathways depending on the node of birth of each protein**

Each color represents a clade. The white rectangles correspond to the genes for which we have not been able to determine the node of birth due to lack of information about the gene. The KEGG legend is available here: [https://www.genome.jp/kegg/document/help\\_pathway.html](https://www.genome.jp/kegg/document/help_pathway.html).

#### **Suppl. Data 4 - Animal tree of life and clades of study**

Tree of life of the 315 species studied here, generated using the information available in Ensembl and Ensembl Metazoa, and with R's ape package (Paradis & Schliep, 2019). The tree is rooted to reflect the phylogeny of the *Opisthokonta*, and the branches are not to scale. The colors are those used in the figures of the article. Each clade is represented by one or more species in our database.

## References

- Albelda, S. M., & Buck, C. A. (1990). Integrins and other cell adhesion molecules. *FASEB Journal: Official Publication of the Federation of American Societies for Experimental Biology*, 4(11), 2868-2880.
- Andreani, J., & Guerois, R. (2014). Evolution of protein interactions : From interactomes to interfaces. *Archives of Biochemistry and Biophysics*, 554, 65-75. <https://doi.org/10.1016/j.abb.2014.05.010>
- Ashton-Beaucage, D., & Therrien, M. (2010). La signalisation RTK/RAS/ERK élargie—Contributions de la génétique à l'assemblage d'un réseau de signalisation. *médecine/sciences*, 26(12), Article 12. <https://doi.org/10.1051/medsci/201026121067>
- Bader, G. D., Cary, M. P., & Sander, C. (2006). Pathguide : A pathway resource list. *Nucleic Acids Research*, 34(Database issue), D504-506. <https://doi.org/10.1093/nar/gkj126>
- Benoit, G., Cooney, A., Giguere, V., Ingraham, H., Lazar, M., Muscat, G., Perlmann, T., Renaud, J.-P., Schwabe, J., Sladek, F., Tsai, M.-J., & Laudet, V. (2006). International Union of Pharmacology. LXVI. Orphan nuclear receptors. *Pharmacological Reviews*, 58(4), 798-836. <https://doi.org/10.1124/pr.58.4.10>
- Berger, M., Stahl, N., Del Sal, G., & Haupt, Y. (2005). Mutations in proline 82 of p53 impair its activation by Pin1 and Chk2 in response to DNA damage. *Molecular and Cellular Biology*, 25(13), 5380-5388. <https://doi.org/10.1128/MCB.25.13.5380-5388.2005>
- Bloom, J. D., Wilke, C. O., Arnold, F. H., & Adami, C. (2004). Stability and the evolvability of function in a model protein. *Biophysical Journal*, 86(5), 2758-2764. [https://doi.org/10.1016/S0006-3495\(04\)74329-5](https://doi.org/10.1016/S0006-3495(04)74329-5)
- Csárdi, G., & Nepusz, T. (2006). *The igraph software package for complex network research*. <https://www.semanticscholar.org/paper/The-igraph-software-package-for-complex-network-Cs%C3%A1rdi-Nepusz/1d2744b83519657f5f2610698a8ddd177ced4f5c>
- Csárdi, G., Nepusz, T., Müller, K., Horvát, S., Traag, V., Zanini, F., & Noom, D. (2023). *igraph for R : R interface of the igraph library for graph theory and network analysis* [Logiciel]. Zenodo. <https://doi.org/10.5281/zenodo.8046777>
- de Vienne, D. M. (2016). Lifemap : Exploring the Entire Tree of Life. *PLOS Biology*, 14(12), e2001624. <https://doi.org/10.1371/journal.pbio.2001624>
- Echave, J., & Wilke, C. O. (2016). *Biophysical models of protein evolution : Understanding the patterns of evolutionary sequence divergence* (p. 072223). bioRxiv. <https://doi.org/10.1101/072223>
- Fraser, H. B., Hirsh, A. E., Wall, D. P., & Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24), 9033-9038. <https://doi.org/10.1073/pnas.0402591101>
- Grandchamp, A., & Monget, P. (2018). Synchronous birth is a dominant pattern in receptor-ligand evolution. *BMC Genomics*, 19. <https://doi.org/10.1186/s12864-018-4977-2>
- Grandchamp, A., & Monget, P. (2020). The membrane receptors that appeared before their ligand : The different proposed scenarios. *PloS One*, 15(5), e0231813. <https://doi.org/10.1371/journal.pone.0231813>
- Howard, A. D., McAllister, G., Feighner, S. D., Liu, Q., Nargund, R. P., Van der Ploeg, L. H., & Patchett, A. A. (2001). Orphan G-protein-coupled receptors and natural ligand discovery. *Trends in Pharmacological Sciences*, 22(3), 132-140. [https://doi.org/10.1016/s0165-6147\(00\)01636-9](https://doi.org/10.1016/s0165-6147(00)01636-9)
- Jack, B. R., Meyer, A. G., Echave, J., & Wilke, C. O. (2016). Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLoS Biology*, 14(5), e1002452. <https://doi.org/10.1371/journal.pbio.1002452>
- Jones, S. N., Roe, A. E., Donehower, L. A., & Bradley, A. (1995). Rescue of embryonic lethality in Mdm2-deficient mice by absence of p53. *Nature*, 378(6553), 206-208. <https://doi.org/10.1038/378206a0>
- Kachroo, A. H., Laurent, J. M., Yellman, C. M., Meyer, A. G., Wilke, C. O., & Marcotte, E. M. (2015). Evolution. Systematic humanization of yeast genes reveals conserved functions and genetic modularity. *Science (New York, N.Y.)*, 348(6237), 921-925. <https://doi.org/10.1126/science.aaa0769>

- Kanki, T., Kurihara, Y., Jin, X., Goda, T., Ono, Y., Aihara, M., Hirota, Y., Saigusa, T., Aoki, Y., Uchiumi, T., & Kang, D. (2013). Casein kinase 2 is essential for mitophagy. *EMBO Reports*, 14(9), 788-794. <https://doi.org/10.1038/embor.2013.114>
- Kauffman, S. A. (1993). *The Origins of Order : Self-organization and Selection in Evolution*. Oxford University Press.
- Kent, R. S., De Lean, A., & Lefkowitz, R. J. (1980). A quantitative analysis of beta-adrenergic receptor interactions : Resolution of high and low affinity states of the receptor by computer modeling of ligand binding data. *Molecular Pharmacology*, 17(1), 14-23.
- Khoury, C. M., & Greenwood, M. T. (2008). The pleiotropic effects of heterologous Bax expression in yeast. *Biochimica Et Biophysica Acta*, 1783(7), 1449-1465. <https://doi.org/10.1016/j.bbamcr.2007.12.013>
- Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences of the United States of America*, 44(2), 98-104. <https://doi.org/10.1073/pnas.44.2.98>
- Kumar, S., Suleski, M., Craig, J. M., Kasprowicz, A. E., Sanderford, M., Li, M., Stecher, G., & Hedges, S. B. (2022). TimeTree 5 : An Expanded Resource for Species Divergence Times. *Molecular Biology and Evolution*, 39(8), msac174. <https://doi.org/10.1093/molbev/msac174>
- Lang, D. T., & Kalibera, T. (2023). *XML : Tools for Parsing and Generating XML Within R and S-Plus* (3.99-0.14) [Logiciel]. <https://cran.r-project.org/web/packages/XML/index.html>
- Lewis, A. C. F., Saeed, R., & Deane, C. M. (2010). Predicting protein-protein interactions in the context of protein evolution. *Molecular bioSystems*, 6(1), 55-64. <https://doi.org/10.1039/b916371a>
- Lovell, S. C., & Robertson, D. L. (2010). An integrated view of molecular coevolution in protein-protein interactions. *Molecular Biology and Evolution*, 27(11), 2567-2575. <https://doi.org/10.1093/molbev/msq144>
- Lynch, M., & Hagner, K. (2015). Evolutionary meandering of intermolecular interactions along the drift barrier. *Proceedings of the National Academy of Sciences*, 112(1). <https://doi.org/10.1073/pnas.1421641112>
- Markov, G. V., Gutierrez-Mazariegos, J., Pitrat, D., Billas, I. M. L., Bonneton, F., Moras, D., Hasserodt, J., Lecointre, G., & Laudet, V. (2017). Origin of an ancient hormone/receptor couple revealed by resurrection of an ancestral estrogen. *Science Advances*, 3(3), e1601778. <https://doi.org/10.1126/sciadv.1601778>
- Maslov, S., & Sneppen, K. (2002). Specificity and stability in topology of protein networks. *Science (New York, N.Y.)*, 296(5569), 910-913. <https://doi.org/10.1126/science.1065103>
- Mintseris, J., & Weng, Z. (2005). Structure, function, and evolution of transient and obligate protein-protein interactions. *Proceedings of the National Academy of Sciences of the United States of America*, 102(31), 10930-10935. <https://doi.org/10.1073/pnas.0502667102>
- Montes de Oca Luna, R., Wagner, D. S., & Lozano, G. (1995). Rescue of early embryonic lethality in mdm2-deficient mice by deletion of p53. *Nature*, 378(6553), 203-206. <https://doi.org/10.1038/378203a0>
- Nguyen, N. T. T., Vincens, P., Dufayard, J. F., Roest Crollius, H., & Louis, A. (2021). Genomicus in 2022 : Comparative tools for thousands of genomes and reconstructed ancestors. *Nucleic Acids Research*, gkab1091. <https://doi.org/10.1093/nar/gkab1091>
- Nguyen, N. T. T., Vincens, P., Roest Crollius, H., & Louis, A. (2018). Genomicus 2018 : Karyotype evolutionary trees and on-the-fly synteny computing. *Nucleic Acids Research*, 46(D1), D816-D822. <https://doi.org/10.1093/nar/gkx1003>
- Nooren, I. M. A., & Thornton, J. M. (2003). Diversity of protein-protein interactions. *The EMBO Journal*, 22(14), 3486-3492. <https://doi.org/10.1093/emboj/cdg359>
- Paradis, E., & Schliep, K. (2019). ape 5.0 : An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, 35(3), 526-528. <https://doi.org/10.1093/bioinformatics/bty633>
- Rand, D. M., Haney, R. A., & Fry, A. J. (2004). Cytonuclear coevolution : The genomics of cooperation. *Trends in Ecology & Evolution*, 19(12), 645-653. <https://doi.org/10.1016/j.tree.2004.10.003>
- Sullivan, S. M., & Holyoak, T. (2008). Enzymes with lid-gated active sites must operate by an induced fit mechanism instead of conformational selection. *Proceedings of the National Academy of Sciences of the United States of America*, 105(37), 13829-13834. <https://doi.org/10.1073/pnas.0805364105>

- Thornton, J. W. (2001). Evolution of vertebrate steroid receptors from an ancestral estrogen receptor by ligand exploitation and serial genome expansions. *Proceedings of the National Academy of Sciences of the United States of America*, 98(10), 5671-5676. <https://doi.org/10.1073/pnas.091553298>
- Tisi, R., Belotti, F., & Martegani, E. (2014). Yeast as a model for Ras signalling. *Methods in Molecular Biology (Clifton, N.J.)*, 1120, 359-390. [https://doi.org/10.1007/978-1-62703-791-4\\_23](https://doi.org/10.1007/978-1-62703-791-4_23)
- Williams, P. D., Pollock, D. D., & Goldstein, R. A. (2001). Evolution of functionality in lattice proteins. *Journal of Molecular Graphics & Modelling*, 19(1), 150-156. [https://doi.org/10.1016/s1093-3263\(00\)00125-x](https://doi.org/10.1016/s1093-3263(00)00125-x)
- Wuchty, S., Oltvai, Z. N., & Barabási, A.-L. (2003). Evolutionary conservation of motif constituents in the yeast protein interaction network. *Nature Genetics*, 35(2), 176-179. <https://doi.org/10.1038/ng1242>
- Zha, H., Fisk, H. A., Yaffe, M. P., Mahajan, N., Herman, B., & Reed, J. C. (1996). Structure-function comparisons of the proapoptotic protein Bax in yeast and mammalian cells. *Molecular and Cellular Biology*, 16(11), 6494-6508. <https://doi.org/10.1128/MCB.16.11.6494>

Figure 1

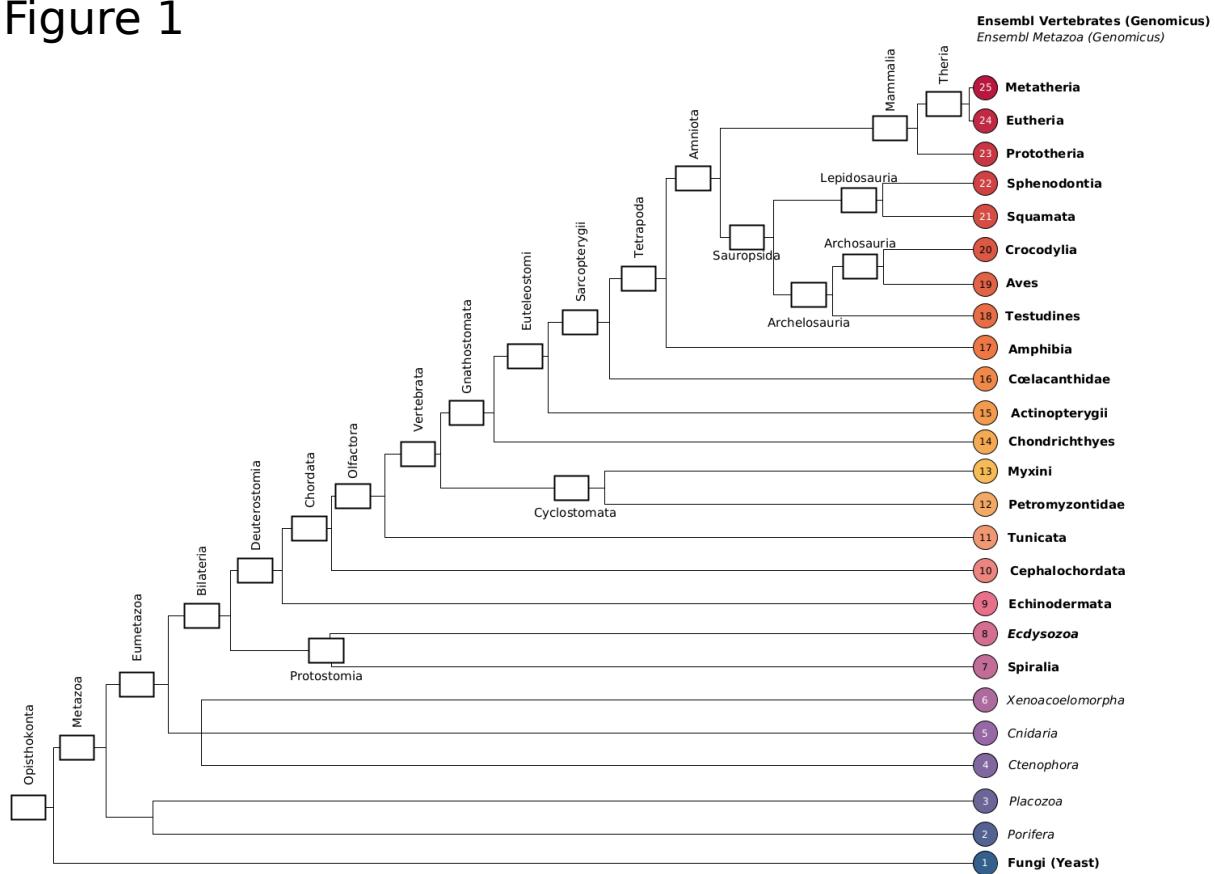


Figure 2

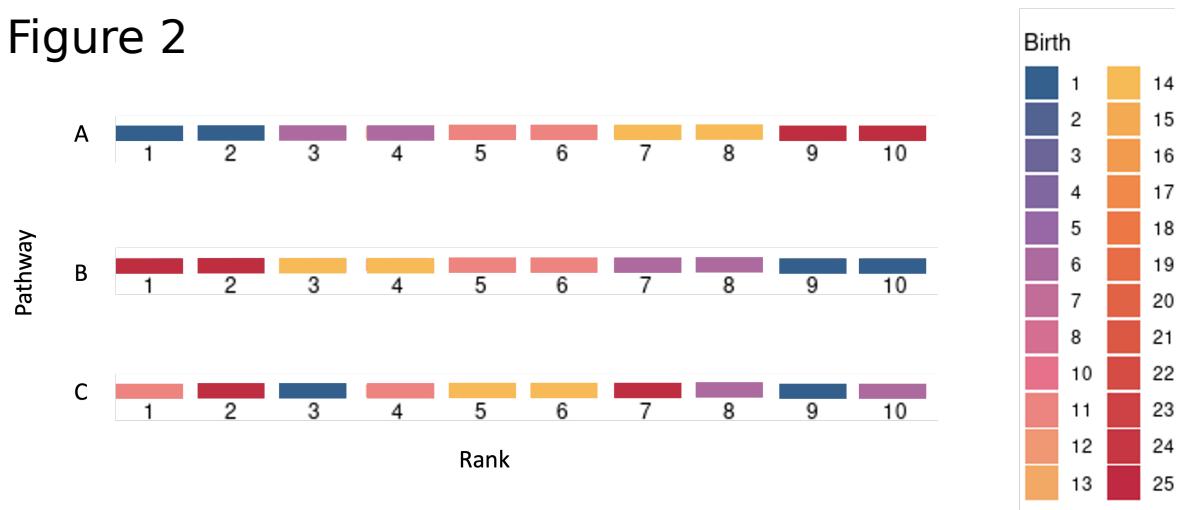


Figure 3

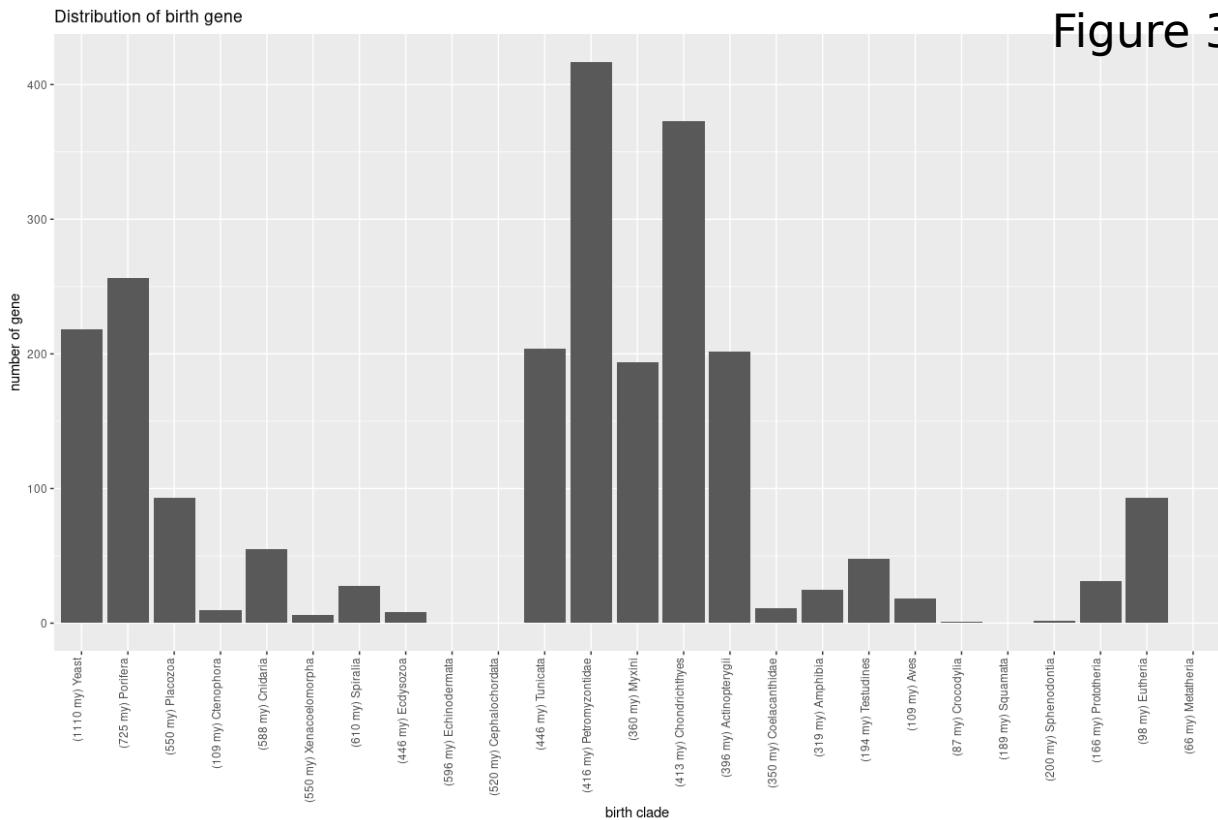
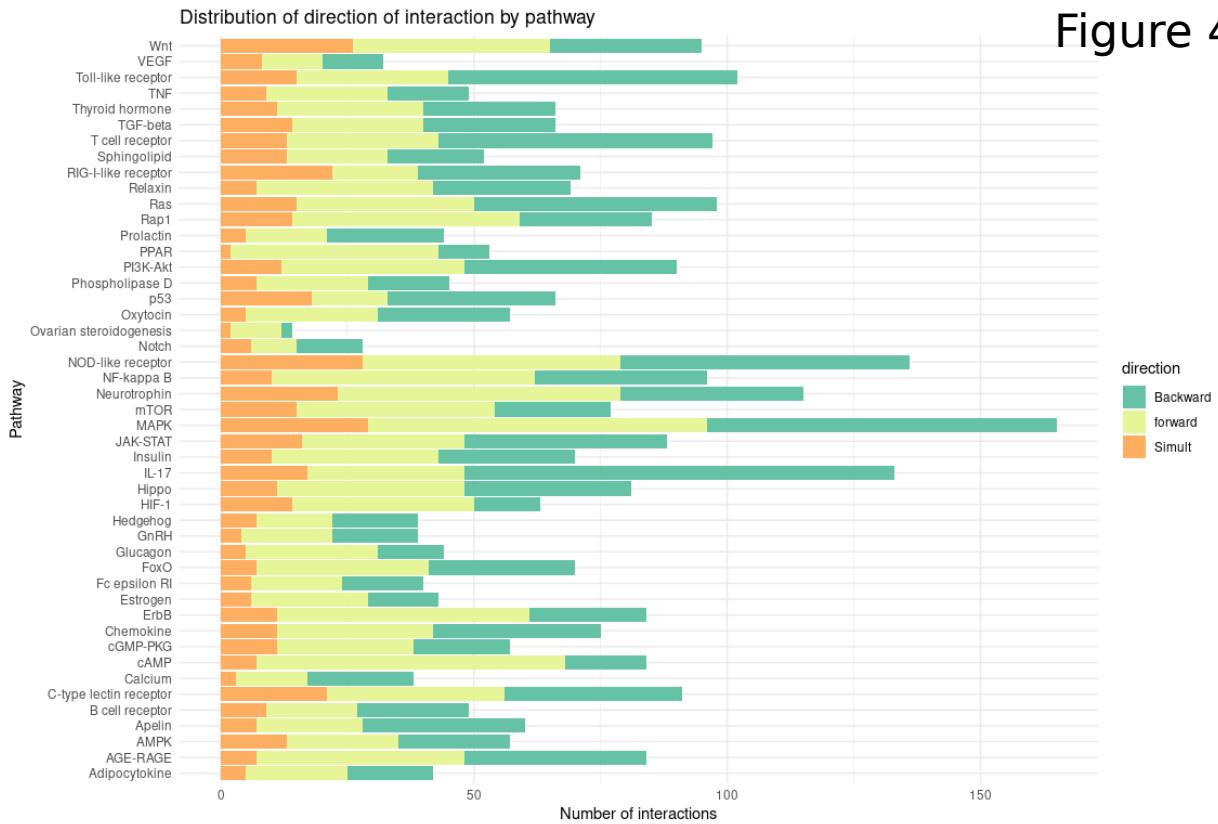


Figure 4



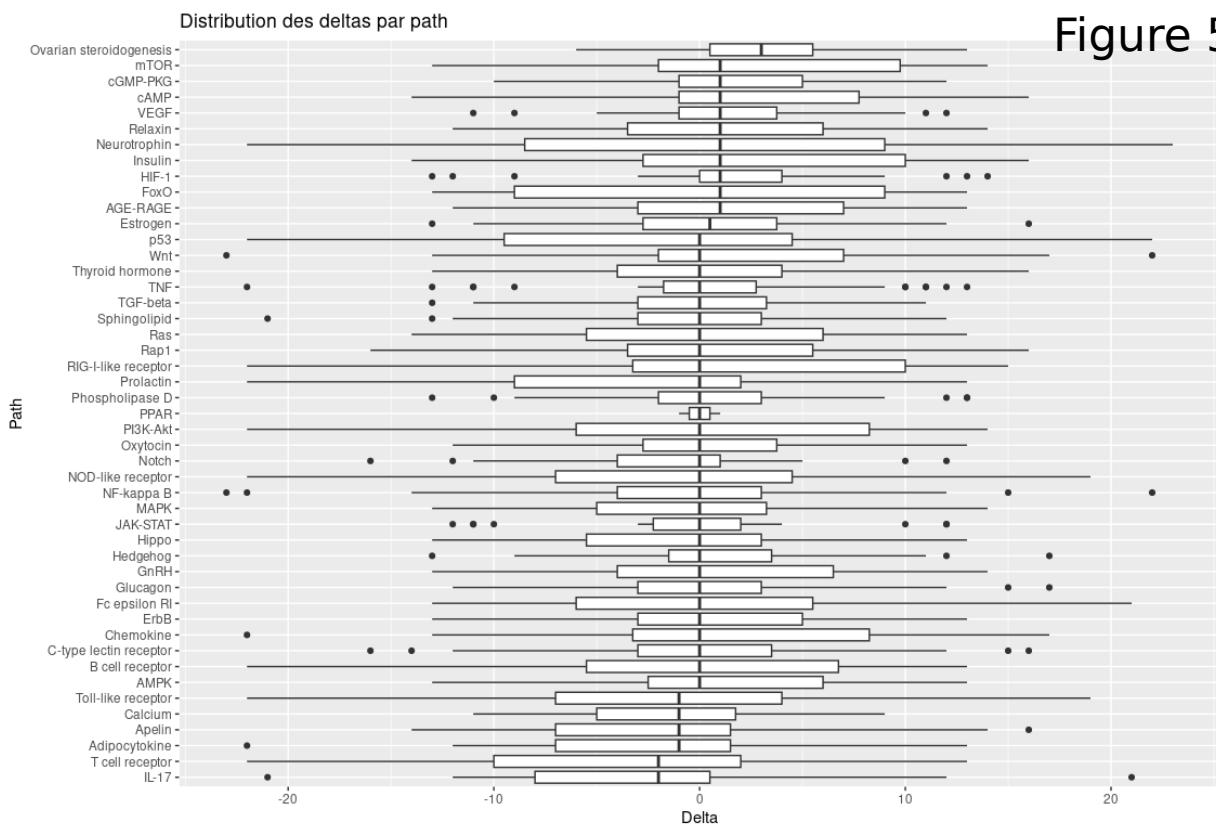


Figure 5

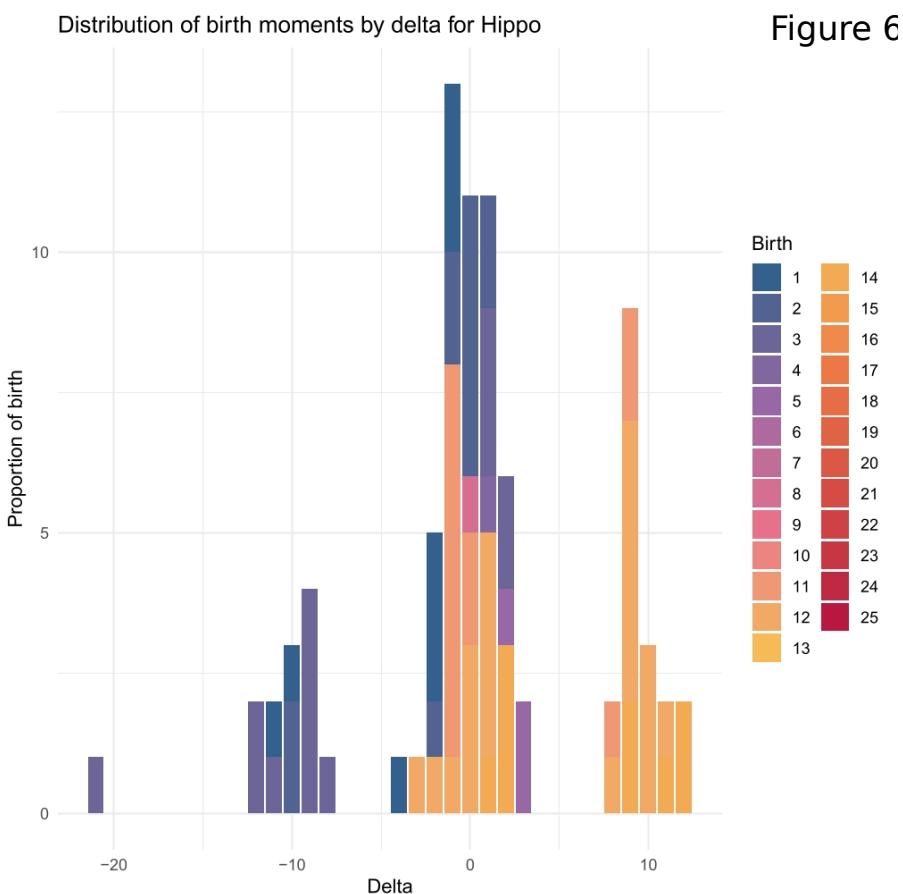


Figure 6

Figure 7

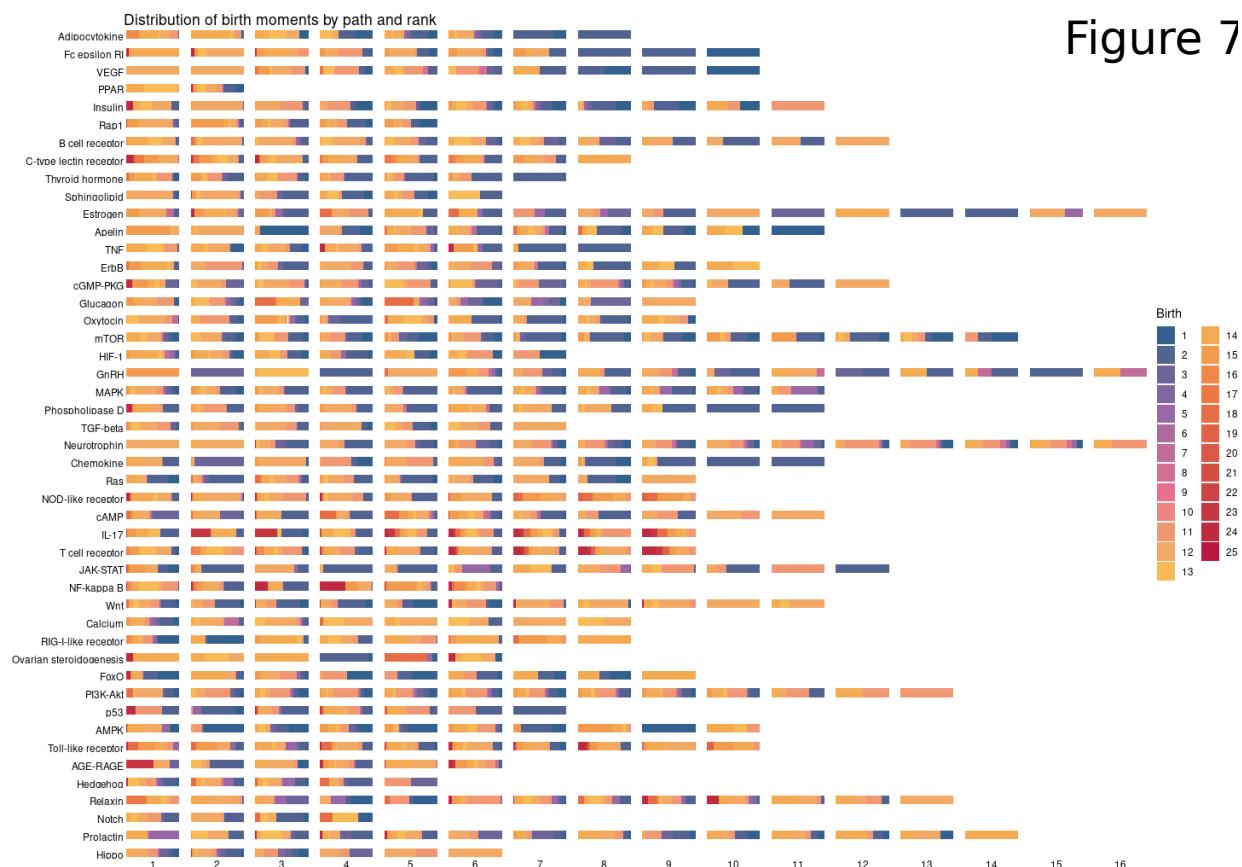
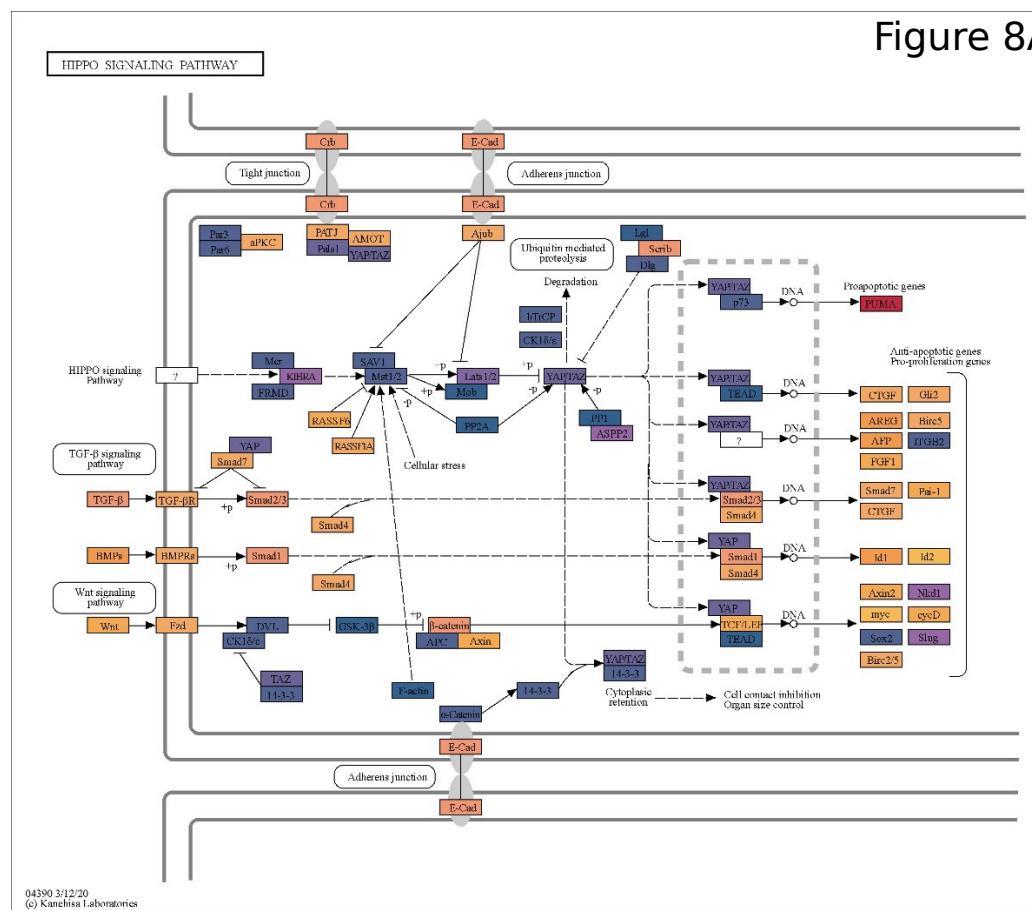
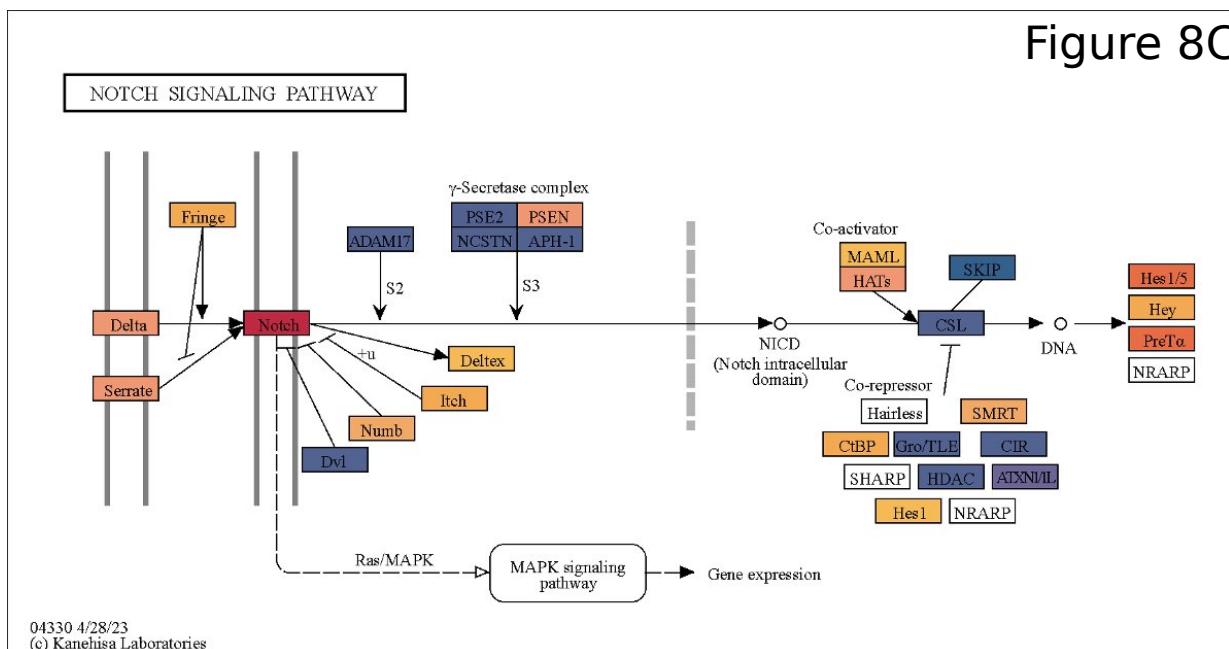
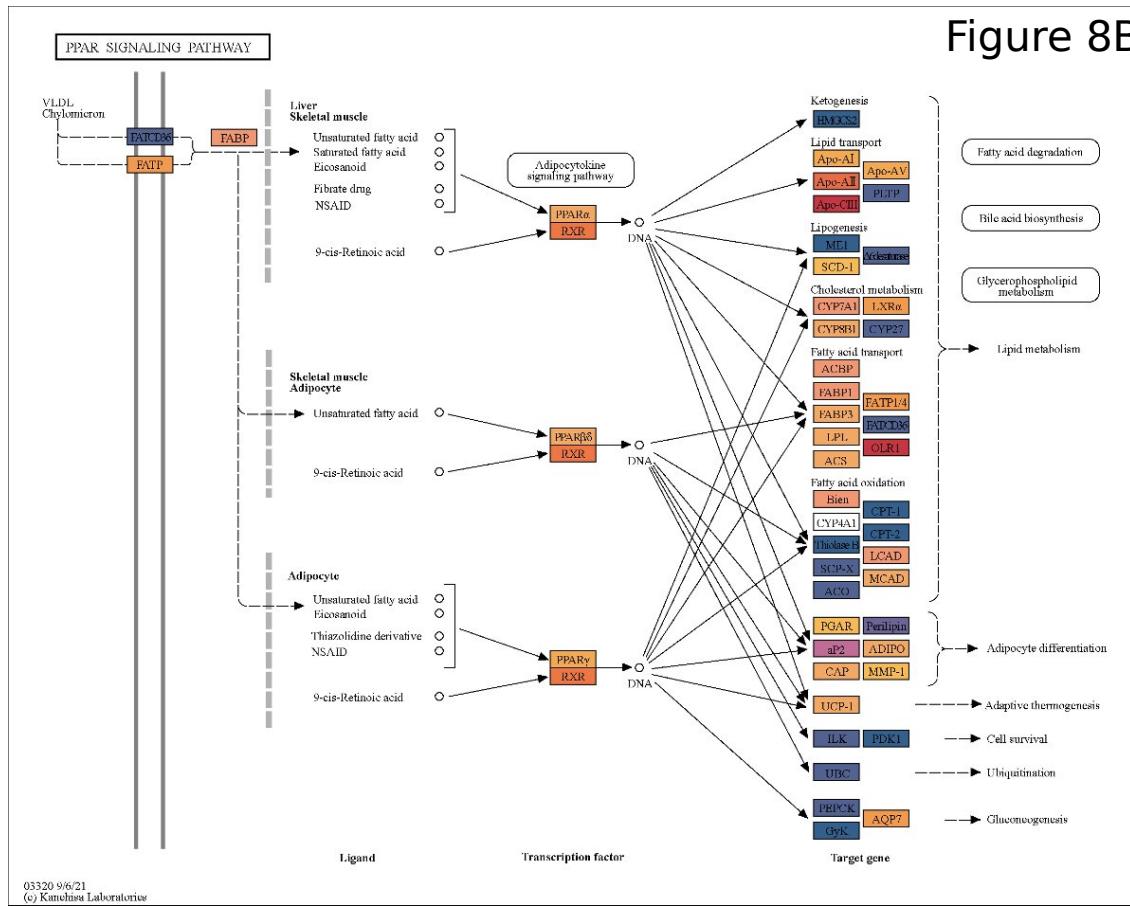
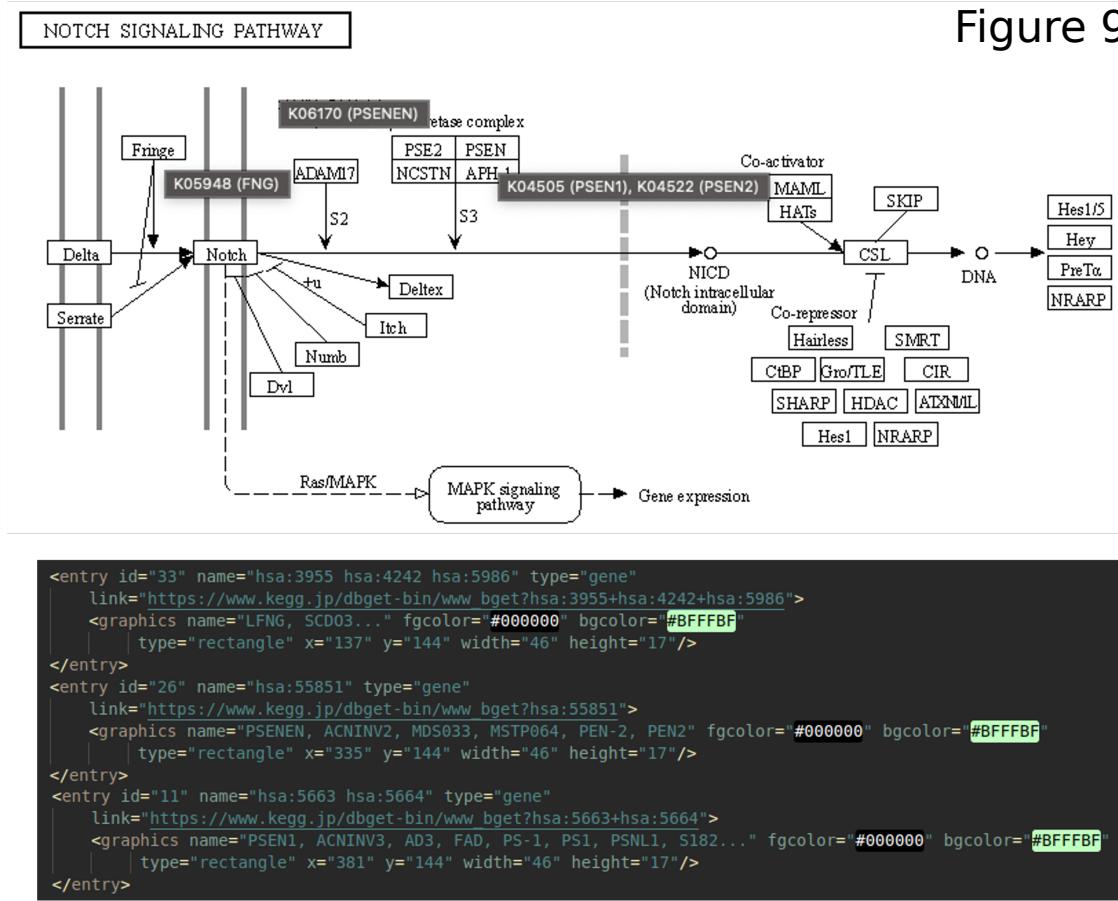


Figure 8A







# Article 2 - Moment d'apparition des gènes impliqués dans une voie de transduction du signal humaine

## 3.1 Contexte de l'étude

Un des piliers majeurs de l'évolution est la naissance de nouveaux gènes, et elle est en partie due aux duplications de gènes et de génomes. Deux duplications de génomes complets (WGD) sont survenues à la racine des vertébrés lors de la scission avec les invertébrés. De plus, le clade des téléostéens a subi une 3ème duplication de génome complet, et une partie de ces espèces (les salmonidés et les carpes) une 4ème. Les périodes de duplications de génome sont suivies de pertes massives de gènes dupliqués. Le clade des téléostéens est donc un modèle de choix pour étudier la capacité de ces gènes à rester à l'état de gènes dupliqués ou leur retour sous forme de singleton. Cette étude est présentée sous forme de schéma récapitulatif en figure ?? page ?? et est suivie d'un article (page ??) publié dans le journal Heliyon.

## 3.2 Matériels et méthode

Pour cette étude, nous avons récupéré un ensemble de 2 298 gènes uniques impliqués dans 47 voies de signalisation humaine dans la base de données KEGG V104. Les voies et leurs caractéristiques sont présentées dans le tableau ??.

Afin de déterminer la quantité d'orthologues téléostéens pour chacun de nos gènes impliqués dans une voie de signalisation, nous avons récupéré l'ensemble des orthologues téléostéens par la plateforme Biomart d'Ensembl V107. L'ensemble des espèces de téléostéens dont le génome annoté est disponible sur Ensembl a été utilisé, soit 63 espèces, dont 54 espèces ayant subi 3 duplications de génome complet, et 9 ayant subi 4 duplications de génome complet. Nous avons conservé 2 groupes distincts pour l'ensemble de l'étude : les

3 WGD, et 4 WGD. Nous avons récupéré toutes les interactions gène-gène pour regarder la proportionnalité des interactions (que nous appellerons « stœchiométries ») dites :

- respectées avec  $n :n$ ,
- non respectées avec  $n :m$  dont  $m > n$ ,
- ou perdues totalement ou partiellement avec  $0 :n$ ,

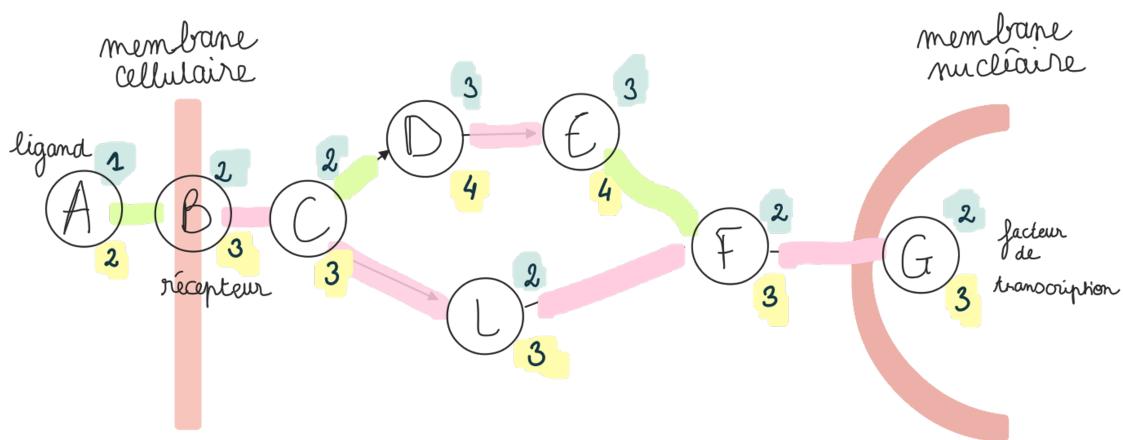
et si une pression subsiste dans la relation pour maintenir les partenaires en quantité égale au sein des espèces, mais également au sein des voies. La proportion des gènes impliqués dans une voie de signalisation dans les différentes stœchiométries sera statistiquement comparée, à l'ensemble des gènes des génomes (Chi2 et test hypergéométrique).

### 3.3 Résultats

Dans un premier temps, nous avons constaté que pour toutes les espèces 3 WGD, nous avons plus souvent retrouvé nos gènes sous la forme dupliquée que les gènes du génome ( $p\text{-value} < 0,001$ ) et pour toutes les espèces 4 WGD, nous avons plus souvent retrouvé nos gènes sous forme triplicat ou plus que les gènes du génome ( $p\text{-value} < 0,00002$ ). Concernant les quantités de gènes pour les interactions gène-gène, nous retrouvons une majorité d'interactions à stœchiométrie 1 :1 (30,3%) pour les espèces 3 WGD, puis les interactions 1 :2 à 27,3% en moyenne et enfin la stœchiométrie 0 :1 à 18% qui représentent une perte partielle de l'interaction chez les téléostéens. Tandis que pour les espèces 4 WGD, on constate une majorité d'interactions 2 :2 (15,8%), puis les interactions 2 :4, à 14% en moyenne et enfin les interactions 0 :2, à 11,6%. Pour les deux groupes étudiés (3 et 4 WGD), nous observons une quantité d'interactions partiellement ou totalement perdues importante (0 :0, 0 :1, 0 :2, 0 :3, 0 :4 allant de 0,1% à 22,1% en fonction des espèces). Les stœchiométries non respectées sont majoritaires dans l'ensemble des voies de signalisation allant de 34% à 70%. Mais 2 voies se démarquent avec une majorité d'interactions  $n :n$  (Hedgehog à 65% et Estrogen à 50%). De plus, aucune interaction n'est totalement perdue pour 3 voies de signalisation chez les téléostéens : JAK-STAT, FoxO et Glucagon.

### 3.4 Conclusion

Nos résultats montrent que les gènes des voies de signalisation restent plus souvent en duplicat ou en triplicat chez les espèces 3 ou 4 WGD de téléostéens que les gènes du génome. Nous retrouvons une majorité d'interactions gène-gène à stœchiométrie respectée en moyenne chez les téléostéens (1 :1 et 2 :2 respectivement pour les 3 et 4 WGD). Cependant, les stœchiométries restent nettement différentes et en fonction de la voie étudiée. Il faut toutefois noter l'exception d'une absence de perte totale pour les 3 voies JAK-STAT, FoxO et Glucagon.



### legende

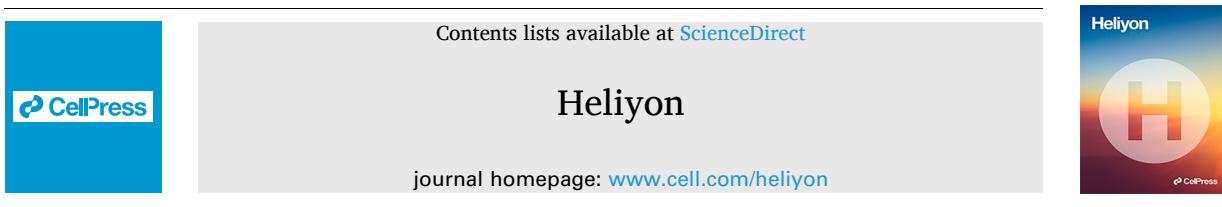
- protéine
- nombre de copie du gène chez une espèce 3 / 4 WGD X- ↗
- interaction directionnelle
- Stoechiométrie respectée
- Stoechiométrie non respectée

### hypothèses

- 1**
- nombre de gènes dupliqués ++ chez X-
- ex: 3 WGD
- 
- |           |   |
|-----------|---|
| singleton | 1 |
| dupliqué  | 3 |
| tripliqué | 2 |
- 2**
- Stoechiométrie respectée ++ chez X-
- ex: 4 WGD
- 
- |     |     |
|-----|-----|
| 2:3 | 4:1 |
| 3:4 | 3:3 |
- 3**
- Stoechiométrie respectée ++ chez les voies

FIGURE 3.1 – Schéma récapitulatif de l'étude 2

### **3.5 Article**



## Genes encoding teleost orthologues of human signal transduction proteins remain duplicated or triplicated more frequently than the whole genome

Floriane Picolo <sup>a</sup>, Benoît Piégu <sup>a</sup>, Philippe Monget <sup>a,\*</sup>

<sup>a</sup> PRC, UMR85, INRAE, CNRS, IFCE, Université de Tours, F-37380 Nouzilly, France



### ABSTRACT

Cell signalling involves a myriad of proteins, many of which belong to families of related proteins, and these proteins display a huge number of interactions. One of the events that has led to the creation of new genes is whole genome duplication (WGD), a phenomenon that has made some major innovations possible. In addition to the two WGDs that happened before gnathostome radiation, teleost genomes underwent one (the 3WGD group) or two (the 4WGD group) extra WGD after separation from the lineage leading to holostei. In the present work, we studied in 63 teleost species whether the orthologues of human genes involved in 47 signalling pathways (HGSP) remain more frequently duplicated, triplicated or in the singleton state compared with the whole genome. We found that these genes have remained duplicated and triplicated more frequently in teleost of the 3WGD and 4WGD groups, respectively. Moreover, by examining pairs of interacting gene products in terms of conserved copy numbers, we found a majority of the 1:1 and 1:2 proportions in the 3WGD group (between 54% and 60%) and of the 2:2 and 2:4 proportions in the 4WGD group (30%). In both groups, we observed the 0:n proportion at a mean of approximately 10%, and we found some pseudogenes in the concerned genomes. Finally, the proportions were very different between the studied pathways. The n:n (i.e. same) proportion concerned 20%–65% of the interactions, depending on the pathways, and the n:m (i.e. different) proportion concerned 34%–70% of the interactions. Among the n:n proportion, the 1:1 ratio is most represented (25.8%) and among the n:m ratios, the 1:2 is most represented (25.0%). We noted the absence of gene loss for the JAK-STAT, FoxO and glucagon pathways. Overall, these results show that the teleost gene orthologues of HGSP remain duplicated (3WGD) and triplicated (4WGD) more frequently than the whole genome, although some genes have been lost, and the proportions have not always been maintained.

### 1. Introduction

Cell signaling involves a large number of proteins, many of which belong to families of more or less related proteins, and these proteins together display a huge number of interactions. One of the events that led to the creation of new genes was whole genome duplication (WGD), which made some major innovations possible. In addition to the two WGDs that happened in vertebrate genomes, the common ancestor of extant teleost fish experienced a third WGD ~320 Million Years Ago (MYA) after separation from the lineage leading to holostei. Moreover, the salmonid and carp lineages have experienced additional independent fourth WGD events 100 and 10 MYA respectively. One of the major assumptions of the preservation of duplicated genes is a dose effect [1]. Indeed, this postulate is based on the idea that the number of copies of a gene is influenced by the dosage of the products with which it interacts (as for example in a signaling pathway) [2]. The products of a signaling pathway should interact in proportion amounts and an imbalance in dosage could lead to a physiological disorder [3]. The functional capacity of a gene lies partly in its dosage, as for example for haplo-insufficient genes which do not produce a functional wild type phenotype if they are in a single copy [4]. We have shown recently that these haplo-insufficient genes are more often maintained in duplicate than the rest of the genes of the genome in teleost

\* Corresponding author.

E-mail address: [philippe.monget@inrae.fr](mailto:philippe.monget@inrae.fr) (P. Monget).

<https://doi.org/10.1016/j.heliyon.2023.e20217>

Received 23 February 2023; Received in revised form 18 August 2023; Accepted 13 September 2023

Available online 16 September 2023

2405-8440/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

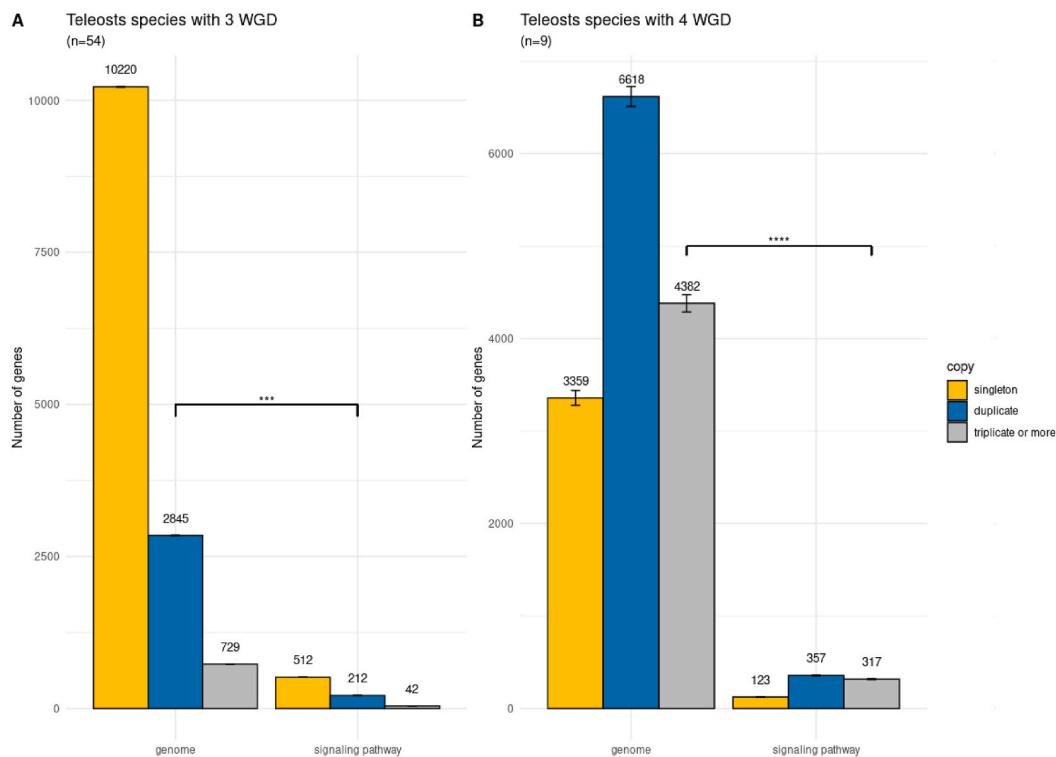
[5]. Most of the interactions between signaling proteins concern enzymatic chain reactions, for which the stoichiometry of the interactions is supposed to be important. From this observation, it would be natural to think that the proportion of proteins involved in these interactions should be of the order of 1:1 (or more generally n:n) for pairs of genes that interact together as may be the case in an intracellular signaling pathway.

The rapid evolution of vertebrates, in terms of diversification and the emergence of species, is notably due to this double WGD, but also due to deletions and more recently species-specific duplications of genes. These processes have rapidly increased the diversity of species [6]. The teleost group, which represents for half of the vertebrates [7], is an interesting group to study from an evolutionary point of view because of its huge phenotypic, environmental and genome size diversity.

The teleost clade (63 teleost species examined in this study) comprises sub-clades that we have divided into two groups in this work. The first sub-clade contains 54 species that have undergone a third WGD (the 3WGD group) after separation from the lineage leading to gars and bowfins (the holostei clade); the second sub-clade includes the salmonid clade (Huchen, Atlantic salmon, Rainbow trout, Coho salmon, Brown trout and Chinook salmon) and the carp clade (Goldfish, Common carp and Golden-line barbel), which have undergone a fourth WGD each (the 4WGD group) [8,9]. Of note, the fourth genome duplication did not occur in the same way in the salmonid and cyprinid groups. The salmonid 4WGD occurred 100 MYA and has been proposed to be an auto-tetraploidization, whereas the cyprinid 4WGD occurred more recently (10 MYA) and has been proposed to be an allo-tetraploidisation [10].

The intracellular signalling pathways are very well referenced in humans, rodents and yeast, but they have been poorly studied in teleosts (there are a little more than 1000 PubMed results with the keywords ‘teleost signalling pathway’ against more than 600,000 PubMed results with the keywords ‘mammals signalling pathway’; <https://pubmed.ncbi.nlm.nih.gov>). Some pathways are also present and referenced in yeast, notably the sphingolipid [11], MAPK [12] or cAMP-PKA [13] signalling pathway. If a pathway is present in yeast, we hypothesise that it is also present in teleosts, even if some orthologues of human genes that encode proteins involved in a signalling pathway (HGSP) may have been lost during evolution.

In the present work, we investigated whether the teleost orthologues of HGSP have remained duplicated (3WGD) or triplicated (4WGD) or returned to the singleton state or in the duplicate state relative to the whole genome.



**Fig. 1.** Bar plot of the global distribution of the genes. (A) The means of teleost 3WGD orthologues, with human genes of the whole genome on the left and HGSP on the right. (B) The means of teleost 4WGD orthologues with human genes of the whole genome on the left and HGSP on the right. The yellow bars correspond to the genes that returned to the singleton state (1 copy); the blue bars correspond to the genes that have been retained (for the 3WGD teleost) or returned (for the 4WGD teleost) to duplicate (two copies). The grey bars correspond to the genes present in three or more copies. The results are presented as the mean  $\pm$  standard error of the mean. \* indicates a significant difference for duplicate orthologues of HGSP ( $***p < 0.001$ ) and triplicate orthologues of HGSP ( $****p < 0.0001$ ) compared with the whole genome of the 3WGD and 4WGD groups, respectively.

## 2. Results

### 2.1. Global distribution of HGSP in 63 teleost species

We found that out of 22,727 human protein-coding genes, on average, 13,875 genes (61.1%) have at least one teleost orthologue. Of these genes, an average of 9240 genes (ranging from 2328 to 10,695 depending on the studied species) have reverted to the singleton state, an average of 3384 genes (ranging from 2423 to 8263) have remained duplicated and an average of 1,251 genes (ranging from 502 to 6374) have three or more copies.

Of the 2989 HGSP, an average of 770 (33.5%) have at least one teleost orthologue. Of these genes, an average of 457 (ranging from 78 to 542) have reverted to the singleton state, an average of 232 (ranging from 178 to 445) have been retained duplicated and an average of 81 genes (ranging from 27 to 422) have three or more copies (Fig. 1 and Suppl. Data 1).

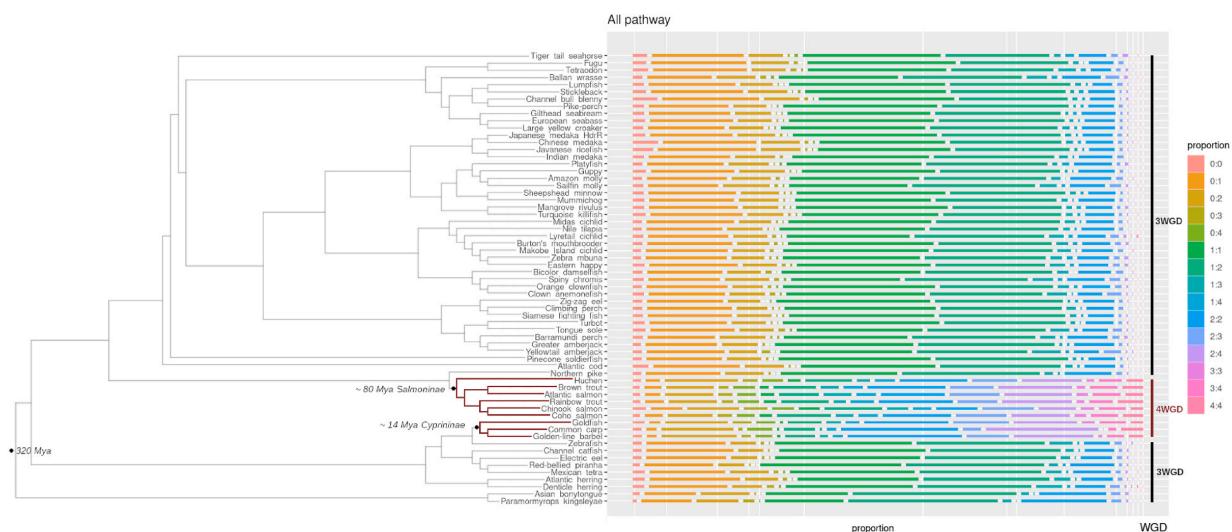
For each of the 54 teleost species in the 3WGD group, more HGSP have been retained in duplicate compared with the whole genome (for 54 species, p-value from 1.00E-03 to 7.74E-15 by  $\chi^2$  analysis after Benjamini–Hochberg [BH] correction, and p-value from 5.48E-04 to 5.49E-04 after hypergeometric test and BH correction, depending on the species considered; see Suppl. Data 1). For each of the nine teleost species in the 4WGD group, more HGSP have been retained in triplicate compared with the whole genome (for all nine species, p-value from 2.00E-05 to 1.33E-08 by  $\chi^2$  analysis after BH correction, and p-value from 2.00E-05 to 9.25E-09 after hypergeometric test and BH correction, depending on the species considered; see Suppl. Data 1). For these 4WGD species, there are not more duplicated HGSP relative to the whole genome, except in the golden-line barrel (BH-corrected p = 1.00E-03).

### 2.2. Global proportion for gene interactions in cellular pathways of each teleost species

In this part of the study, we did not include the PPAR pathway because in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database, it is only represented by gene expression, instead of protein–protein interactions. The signalling pathways are characterised by a chain of protein–protein interactions, which led us to consider whether the proportion of gene–gene interactions (in our case) is respected. First, we determined whether the proportion was maintained in some teleost species and not in others, considering both the 3WGD and 4WGD groups.

For the 3WGD group (Fig. 2), the 1:1 proportion was the highest: a mean of 30.3% with a range from 17.7% (*Paramormyrops kingsleyae*) to 35.2% (Turbot). We observed the 2:2 proportion at a mean of 7.2%, ranging from 4.4% (Lumpfish) to 19.2% (*P. kingsleyae*); the 3:3 proportion had a mean of 0.1%. The 1:2 proportion had a mean of 27.3% with a range from 23.0% (*Tetraodon*) to 36.6% (Asian bonytongue). Finally, we observed the 0:1 proportion at a mean of 18.0% with a range from 11.2% (*P. kingsleyae*) to 22.1% (Javanese ricefish).

For the 4WGD group (Fig. 2), we found a lower 0:n proportion than in the 3WGD group. We found a mean of 2.0% for the 1:1 interaction. The 2:2 proportion was the highest in the 4WGD group: a mean of 15.8% with a range from 7.7% (Chinook salmon) to 24.9% (Common carp). The 2:4 proportion had a mean of 14.0% with a range from 6.9% (Chinook salmon) to 23.9% (Common carp). For the 0:2 proportion, we observed a mean of 11.6% with a range from 9.7% (Chinook salmon) to 15.3% (Common carp) (Fig. 2). This third proportion represents the loss of one of the elements of the interaction while the other partner has returned to being present in



**Fig. 2.** Distribution of the gene–gene interaction proportions by teleost species for all genes involved in signalling pathways. The tree of life of teleosts is presented on the left side. The nodes with a bullet represent a WGD. The black branches are the species with 3WGD, and the red branches are the species with 4WGD. The tree is not to scale, and the genome duplication times have been added according to previous studies [9,22–24]. The right side shows the distribution of gene–gene interaction proportions. Each bar colour represents one proportion.

duplicate. In the 4WGD group, the Common carp showed the highest 2:4 (23%) and 2:2 (24%) proportions, while the Chinook salmon showed the lowest percentages (7%). Interestingly, despite these different ‘histories’ of fourth duplication between the salmonid and cyprinid groups, the proportions between partner proteins are very similar between both groups (Fig. 2).

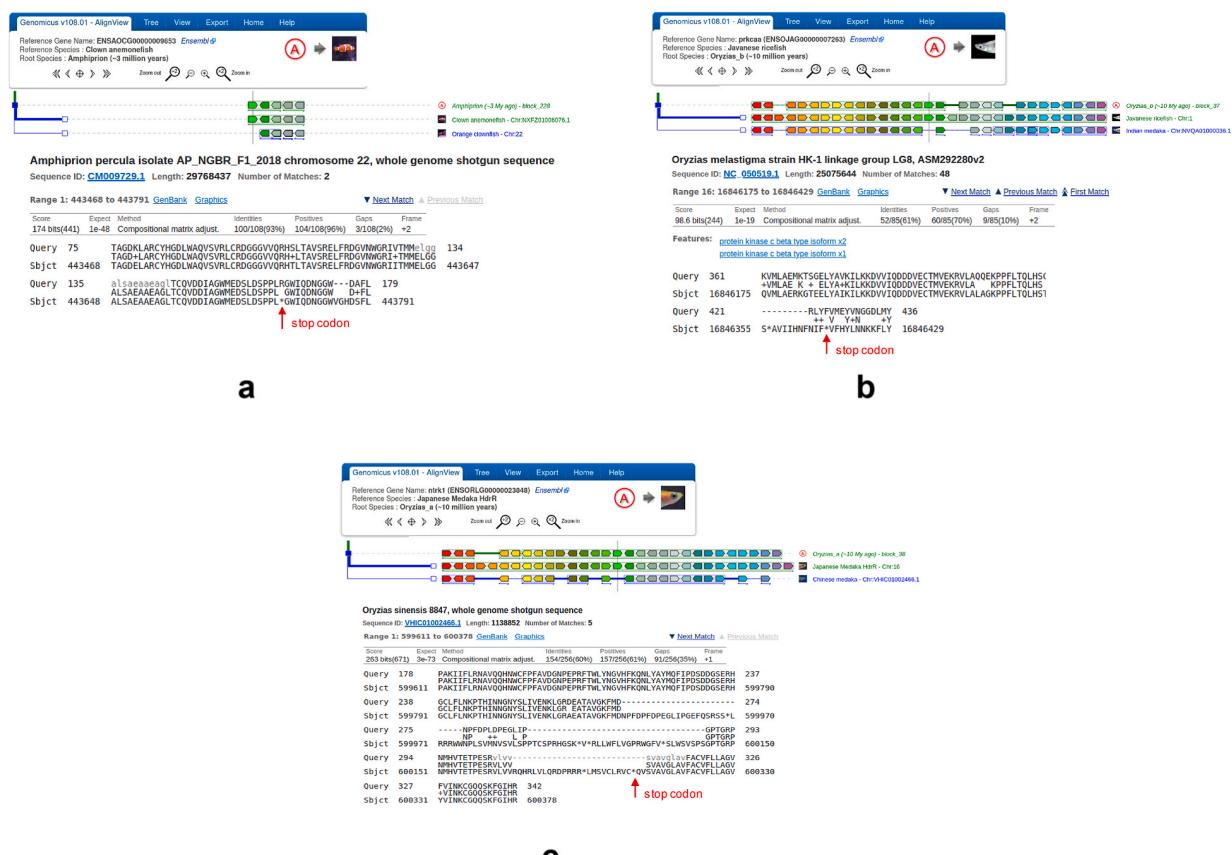
In both the 3WGD and 4WGD groups, a very large majority of genes almost never respect an n:n proportion (in more than 95% of the species studied here). These genes belong to all signaling pathways.

For both the 3WGD and 4WGD groups, the 0:0, 0:1, 0:2, 0:3 and 0:4 proportions occurred at a mean of 5.8% with a range from 0.1% to 22.1%. We observed the complete loss of interaction – with the loss of both partners, *PYCARD* and *CASP8* (corresponding to the 0:0 proportion) – for all the species except *P. kingsleyae*, which diverged early from the lineage that contains the majority of teleost species in the tree of life. Moreover, several genes are absent in a large majority of the 63 studied species except for less than three species (they are never the same species), and for which the proportion had not been maintained. Once again, the pycard gene is absent in a majority of species and the other concerned genes encode proteins involved in different pathways: *bad*, *ccl26*, *bid*, *map2k3*, *traf3ip2*, *casp8*, *nlrp1*, *nfbka*, *csnk2a1*, *cd40lg*, *mefv*, *bbc3*, *prkd3*, *rap1a*, *tp53aip1*, *ticam1*, *traf3*, *ywhaq*, *tir4*, *nlrp12*, *ly96*, *nlrp3*, *ifi16*, *aim2*, *icos*, *rgs14*, *cycs*, *cxcr6*, *nlrp6*, *pydc2* and *pydc5*, some of which interact with pycard (see Suppl. Data 2).

These 0:n proportions strongly suggest the loss by pseudogenisation of several teleost orthologues of HGSP. We failed to find the pycard pseudogenes in the 62 teleosts in which the gene is absent from the Ensembl phylogenetic trees in the teleost except in *P. kingsleyae*, probably because the evolutionary distance is too great. On the other hand, we found some other pseudogenes like that of *bcl2 like*, present in two copies in Clown anemonefish and only in one copy in Orange clownfish. We also found a pseudogene *prkca* in the Indian medaka species (Fig. 3).

### 2.3. Proportion of gene interactions in all teleost species for each cellular pathway

We evaluated the proportion of gene–gene interactions for each of the intracellular signalling pathways. We classified the different proportions into the following categories: 0:0; n:n, representing equal proportions (1:1, 2:2, 3:3 and 4:4); and n:m, representing



**Fig. 3.** Pseudogenes found by tblastn. (A) A screen capture of Genomicus (<https://www.genomicus.bio.ens.psl.eu/genomicus-100.01/cgi-bin/search.pl>), which we used to visualise a gene present in one species and absent in another. The gene of interest is in light green. We chose species that are close to each other to maximise our chances to find a trace of our genes. (B) Protein sequence alignment of the gene of the closest species against our species of interest. This is a tblastn alignment against the whole genome of the target species. The stars represent the stop codons in the sequence.

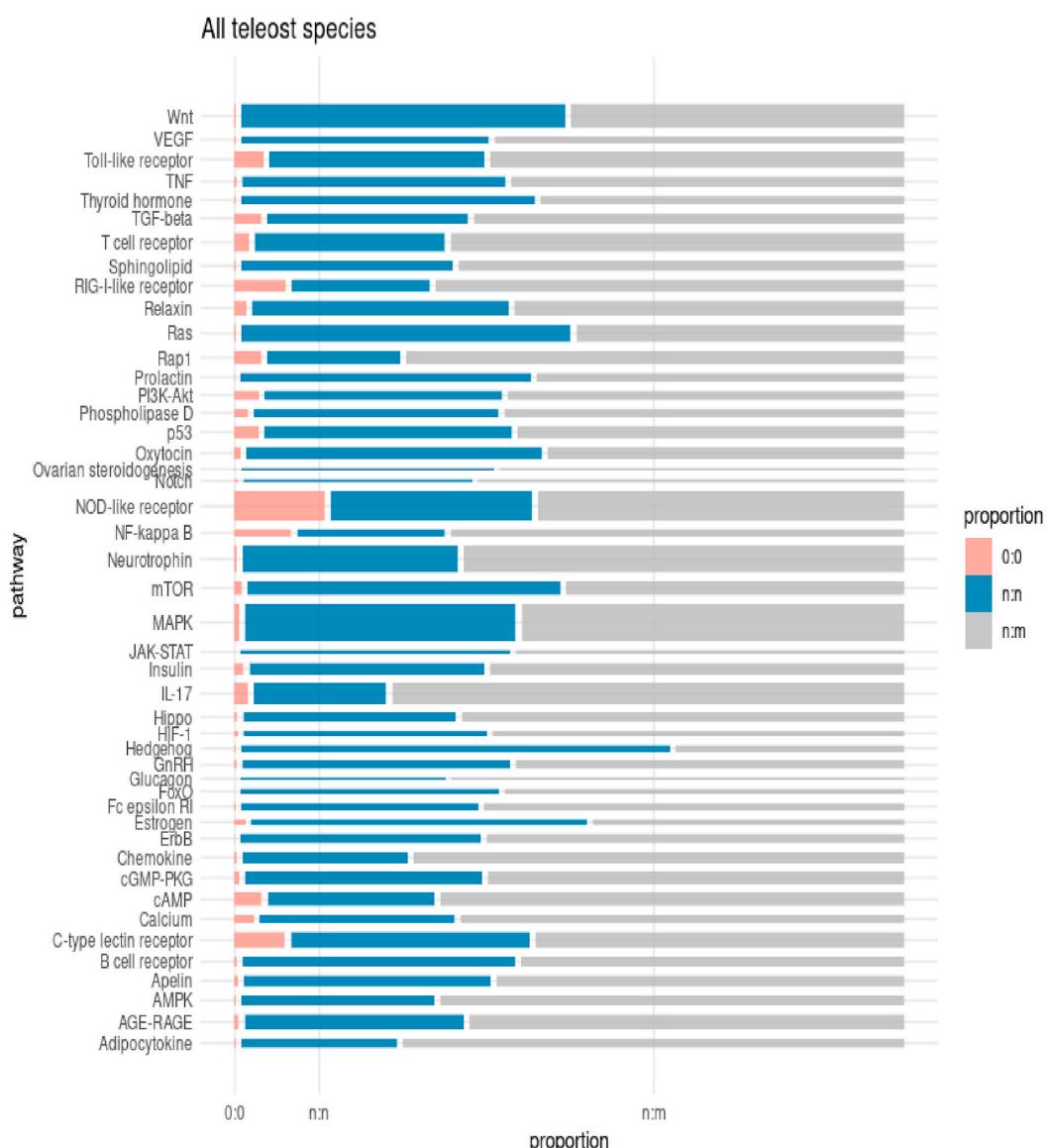
proportions that have not been maintained (0:1, 0:2, 1:2, ..., 3:4). We found that the proportion of gene–gene interactions is very different between the pathways, with the n:n proportion being more represented in the Hedgehog and oestrogen pathways than the *Rap1*, *IL-17* or *RIG-I-like* receptor pathways. The percentage of this type of interaction decreased among all pathways, from 65% to 20%. Among the n:n proportions, the 1:1 proportion had a mean of 71.4% of interactions, the 2:2 had a mean of 26.0% and the 3:3 and 4:4 proportions had a mean of 0.9% and 1.7%, respectively (Fig. 4).

The n:m proportion is widely represented in all signalling pathways, ranging from 34% (Hedgehog) to about 70% (*Rap1*, *IL-17*, *RIG-I-like* receptor and adipocytokine). The proportions most represented in the n:m category are 1:2 (mean 44.3%), 0:1 (mean 24.2%), 0:2 (mean 13.2%) and then the other proportions (mean 18.3%).

We observed no gene loss for the *JAK-STAT*, *FoxO* and glucagon pathways. Most gene loss (>7.7%) occurred for the C-type lectin receptor, *RIG-I-like* receptor, *NOD-like* receptor and *NF-kappa B* pathways (see Suppl. Data 3).

### 3. Discussion

We found that the teleost orthologues of HGSP remained duplicated for the 3WGD species and triplicated for the 4WGD species



**Fig. 4.** Distribution of orthologues of the human gene–gene interaction proportions by signalling pathway species for all genes with an HGSP orthologue in teleost species. Each bar colour represents one type of proportion: 0:0 (pink), n:n (1:1, 2:2, 3:3 and 4:4, blue) and n:m (0:1, 0:2, ..., 3:4, grey). The thickness of each bar is proportional to the number of unique interactions in each pathway.

more frequently than the whole genome. We obtained similar results for teleost orthologues of genes encoding ligand/membrane receptor pairs [14]. This previous work had only been carried out with 10 teleost species. Here we have studied 63 species, which strengthens our conclusions. The fact that these genes remained in duplicate or more instead of returning to singleton suggest that there is selection pressure that keeps them in duplicate, triplicate or more. The hypothesis is that evolution would favor the maintenance of these HGSP genes in large quantities because they would thus be more favorable to the survival of the species.

Concerning signalling proteins, as for ligand–receptor pairs, one of the questions raised here concerns whether the proportion of the interactions studied has been respected. In the case of ligand–receptor interactions, partners have not necessarily evolved in the same way, and there have been situations in which one of the partners has returned to the singleton state while the other one has been maintained in duplicate [14]. This suggests that changes in ligand–receptor interactions may have taken place during the evolution of teleosts. In the present study, the n:n proportion was more represented in the 3WGD group (>37.5%) than in the 4WGD group (>23.1%). Interestingly, among the genes that remained duplicated or triplicated depending on the group of teleost concerned, we found 74 teleost gene orthologues of HGSP that are haplo-insufficient (*APC*, a regulator of the *WNT* signalling pathway; *AMT* [serine/threonine kinase]; *JAG1* [jagged canonical Notch ligand 1]; and 107 orthologues of HGSP that are mono-allelically expressed [*cd38*, *cd86*, *cd72*, etc.]). We have previously studied these genes [5], a fact that strengthens the conclusions of the present work. Interestingly, despite these different 4WGD ‘histories’ between the salmonid and cyprinid groups, the proportions between partner proteins are very similar in both groups (Fig. 2). This suggests that the process of polyploidisation does not influence whether the genes encoding teleost HGSP return to triplicates or remain in quadrupletes.

Only one interaction, *CASP1:CARD16*, respects the strict proportion (from 2:2 in *Tetraodon* to 14:14 in *Zebra mbuna*) among the 63 teleost species we studied. This finding strongly suggests evolutionary pressure to maintain a strict equilibrium in term of the concentration of both partners. Several genes respect this n:n proportion in at least 95% of teleost species. The concerned genes encode proteins involved in different pathways: *CASP8*, *NOS1*, *MAPK1*, *PRKCA*, *TP53*, *BAX*, *FZD10*, *DVL1*, *FOS*, *GNA12*, *CALM6*, *SRC*, *NFkB* and *MTOR*, among others. This suggests that the n:n proportion of these partners has been maintained during the evolution of certain species, but it has been relaxed in others. However, we do not know whether these paralogues are expressed in the same cells or if some have not become sub-functionalised or neo-functionalised.

The presence of several cases of 0:n interactions strongly suggests that several orthologues of HGSP have been lost during evolution. Although we found some pseudogenes (bcl2-like, nerve growth factor receptor), we failed to find many others. It is possible that the evolutionary distance is too great to find any stop codon or other frameshift mutations, insertions or deletions. It is also possible that some of these predicted absent genes are truly present but have not been sequenced or annotated in these recently sequenced genomes. Moreover, there are about 74,962 protein-coding genes in the genome of 4WGD species (range from 55,255 to 99,592), and about 33,640 protein-coding genes in the genome of 3WGD species range from 23,886 to 100,231). Regarding the HGSP orthologues, there are about 14,359 genes in the 4WGD species, and 13,794 in the 3WGD species. This strongly suggests that the genomes of the 4WGD group have lost many more HGSP orthologues than the genomes of the 3WGD group [10], likely due to a more intense process of pseudogenization.

Regarding these predicted 0:n interactions, it is unlikely that the signalling pathway functions without the presence of one of its members. On the other hand, it is very probable that the lost gene has been replaced by a parologue of the same family which could replace the member described in KEGG. In particular, the majority if not all of these genes belong to families (bad, mapk, casp, etc.) for which functional redundancy is well documented [14,15].

Some genes have unusual characteristics. For example, *clec4d*, *clec4m* and *clec6a* are present in 3–33 copies in the 63 teleosts we studied. These genes encode membrane receptors with an extracellular C-type lectin-like domain that recognises several pathogens, and are involved in inflammation and immune response, suggesting a particularly important role for these biological functions in teleost species. The *stat1* gene is also present in 15 copies in Ballan wrasse; the *casp1* gene is present in 10, 14 and 12 copies in Eastern happy, Zebra mbuna and Golden-line barbel; and the *IFNG:IFNCR1* interaction is present as a 4:13, 4:16 and 8:17 ratio in Golden-line barbel, Common carp and Goldfish, respectively. Additional investigations are needed to understand the biological significance of these duplications.

The *p53* gene also caught our attention: there are 25 copies in the Siamese fighting fish, reminiscent of the massive duplications of the same gene in the elephant [15,16]. In this latter species, it has been suggested that this duplication is partly responsible for the low incidence of cancer and its long life expectancy (Peto’s paradox). The Siamese fighting fish is an aquarium fish with a size of 6–8 cm and a life expectancy that is not particularly long (3–5 years in captivity). It is possible that this teleost develops cancer less frequently than other aquarium fish species. However, it is probably difficult to compare the elephant and an aquarium fish in terms of life expectancy and cancer incidence.

Overall, our work shows that teleost genes orthologues of HGSP remain more duplicated in the 3WGD group and more triplicated in the 4WGD group compared with the whole genome. Moreover, some genes involved in almost all pathways studied have been lost, and there are many protein–protein interactions for which the proportion has not been maintained.

#### 4. Material and methods

##### 4.1. Implementation of the database

We focused on human genes coding for protein involved in 47 different signalling pathways. We obtained the human genes of these pathways with KEGG database V104.0 (<https://www.genome.jp/kegg>), which is the most popular metabolic database [17], by using the keywords ‘signalling pathway’ or ‘ovarian’ and ‘human’. These signalling pathways are named as follows in KEGG: *PPAR*, *MAPK*,

*ErbB, Ras, Rap1, Calcium, cGMP-PKG, cAMP, Chemokine, NF-kappa B, HIF- 1, FoxO, Sphingolipid, Phospholipase D, p53, mTOR, PI3K-Akt, AMPK, Wnt, Notch, Hedgehog, TGF-beta, VEGF, Apelin, Hippo, Toll-like receptor, NOD-like receptor, RIG-I-like receptor, C-type lectin receptor, JAK-STAT, IL-17, T cell receptor, B cell receptor, Fc epsilon RI, TNF, Neurotrophin, Insulin, GnRH, Ovarian steroidogenesis, Estrogen, Prolactin, Thyroid hormone, Adipocytokine, Oxytocin, Glucagon, Relaxin and AGE-RAGE.*

The number of genes encoding proteins involved in these signalling pathways varies from 51 to 353, and several proteins are involved in different pathways. We investigated a total of 2295 genes (see Suppl. Data 2).

We studied 63 species of teleosts available on Ensembl database: Amazon molly (*Poecilia formosa*), Asian bonytongue (*Scleropages formosus*), Atlantic cod (*Gadus morhua*), Atlantic herring (*Clupea harengus*), Atlantic salmon (*Salmo salar*), Ballan wrasse (*Labrus bergylta*), Barramundi perch (*Lates calcarifer*), Bicolour damselfish (*Stegastes partitus*), Brown trout (*Salmo trutta*), Burton's mouthbrooder (*Haplochromis burtoni*), Channel bull blenny (*Cottoperca gobio*), Channel catfish (*Ictalurus punctatus*), Chinese medaka (*Oryzias sinensis*), Chinook salmon (*Oncorhynchus tshawytscha*), Climbing perch (*Anabas testudineus*), Clown anemonefish (*Amphiprion ocellaris*), Coho salmon (*Oncorhynchus kisutch*), Common carp (*Cyprinus carpio*), Denticle herring (*Denticeps clupeoides*), Eastern happy (*Astatotilapia calliptera*), Electric eel (*Electrophorus electricus*), European seabass (*Dicentrarchus labrax*), Fugu (*Takifugu rubripes*), Gilthead seabream (*Sparus aurata*), Golden-line barbel (*Sinocyclocheilus grahami*), Goldfish (*Carassius auratus*), Greater amberjack (*Seriola dumerili*), Guppy (*Poecilia reticulata*), Huchen (*Hucho hucho*), Indian medaka (*Oryzias melastigma*), Japanese medaka HdrR (*Oryzias latipes*), Javanese ricefish (*Oryzias javanicus*), Large yellow croaker (*Larimichthys crocea*), Lumpfish (*Cyclopterus lumpus*), Lyretail cichlid (*Neolamprologus brichardi*), Makobe Island cichlid (*Pundamilia nyererei*), Mangrove rivulus (*Kryptolebias marmoratus*), Mexican tetra (*Astyanax mexicanus*), Midas cichlid (*Amphilophus citrinellus*), Mummichog (*Fundulus heteroclitus*), Nile tilapia (*Oreochromis niloticus*), Northern pike (*Esox lucius*), Orange clownfish (*Amphiprion percula*), P. kingsleyae, Pike-perch (*Sander lucioperca*), Pinecone soldierfish (*Myripristis murdjan*), Platypfish (*Xiphophorus maculatus*), Rainbow trout (*Oncorhynchus mykiss*), Red-bellied piranha (*Pygocentrus nattereri*), Sailfin molly (*Poecilia latipinna*), Sheepshead minnow (*Cyprinodon variegatus*), Siamese fighting fish (*Betta splendens*), Spiny chromis (*Acanthochromis polyacanthus*), Stickleback (*Gasterosteus aculeatus*), Tetraodon (*Tetraodon nigroviridis*), Tiger tail seahorse (*Hippocampus comes*), Tongue sole (*Cynoglossus semilaevis*), Turbot (*Scophthalmus maximus*), Turquoise killifish (*Nothobranchius furzeri*), Yellowtail amberjack (*Seriola lalandi dorsalis*), Zebra mbuna (*Maylandia zebra*), Zebrafish (*Danio rerio*) and Zig-zag eel (*Mastacembelus armatus*).

The genomes of these teleost species have been subjected to a third duplication after divergence with tetrapod (the 3WGD group) or to a fourth duplication (the 4WGD group). The human genes were extracted from Ensembl (<http://www.ensembl.org>) [18].

## 5. Analysis

We generated a list of human genes (GRCh38.p13) by using BioMart from Ensembl Genes 107. We selected the set of human genes coding for a protein (protein coding) in the gene type filter. The attributes selected in the homologous category were the different teleost species listed in Ensembl. We only selected stable gene identifiers. We generated a list of 22,881 human protein-coding genes. We established the orthologue of each gene in each of the 63 fish species. Then, in each species, we studied the fate (loss of gene, singleton, duplicate, triplicate or more) of all HGSP orthologues. We obtained between 12,863 (Tetraodon) and 14,542 (Brown trout) orthologous genes per fish species (mean 13,878). This does not represent the entire genome of each fish, but it allowed for solid statistical analysis. We investigated whether these fish orthologues of signalling transduction genes remained in triplicate, in duplicate or returned to the singleton state relative to the whole genome. For the statistical analysis, we used the  $\chi^2$  test and a hypergeometric analysis with BH correction to test these hypotheses. Regarding genes encoding proteins of signalling pathways that interact in pairs, we recovered a total of 2317 single gene–gene interactions among the 47 pathways of the KEGG database. We used R for all statistical analysis.

## 6. Identification of pseudogenes

We also performed a systematic search for pseudogenes by using tblastn in the studied genomes for genes with no orthologue identified in at least one of the species of interest. This allowed us to test the hypothesis that evolution of teleost orthologues of HGSP is partly characterised by a gene loss pattern, as previously described for seminal plasma genes and for genes encoding proteins of oviductal fluids in mammals [19–21]. We inferred the pseudogene status in a genome if we found a stop codon or an indel in the sequence identified by the similarity search in the syntenic locus in comparison with the other species of interest. For genes that we did not find, and for which there was no pseudogene in the syntenic locus, we can only hypothesise that the gene has been lost.

## Author contribution statement

Floriane Picolo: Performed the experiments; Analyzed and interpreted the data; Wrote the paper. </p>

Benoît Piégu: Analyzed and interpreted the data. </p>

Philippe Monget: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper. </p>

## Data availability statement

The authors are unable or have chosen not to specify which data has been used.

### Declaration of competing interest

I hereby declare that the disclosed information is correct and that no other situation of real, potential or apparent conflict of interest is known to me. I undertake to inform you of any change in these circumstances, including if an issue arises during the course of the meeting or work itself.

### Acknowledgements

The authors are grateful to Dr Julien Bobe and Yann Guiguen for helpful discussions. Furthermore, the authors are very grateful to Alexandra Louis for discussion and their phylogenetic trees.

### Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.heliyon.2023.e20217>.

### References

- [1] B. Papp, C. Pál, L.D. Hurst, Dosage sensitivity and the evolution of gene families in yeast, *Nature* 424 (6945) (2003), <https://doi.org/10.1038/nature01771>. Art. n° 6945, juill.
- [2] M. Freeling, B.C. Thomas, Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity, *Genome Res.* 16 (7) (2006) 805–814, <https://doi.org/10.1101/gr.3681406>, juill.
- [3] T. Makino, A. McLysaght, Ohnologs in the human genome are dosage balanced and frequently associated with disease, *Proc. Natl. Acad. Sci.* 107 (20) (2010) 9270–9274, <https://doi.org/10.1073/pnas.0914697107>, mai.
- [4] V.T. Dang, K.S. Kassahn, A.E. Marcos, M.A. Ragan, Identification of human haploinsufficient genes and their genomic proximity to segmental duplications, *Eur. J. Hum. Genet.* 16 (11) (2008) 1350–1357, <https://doi.org/10.1038/ejhg.2008.111>, juin.
- [5] F. Picolo, A. Grandchamp, B. Piégut, A.D. Rolland, R.A. Veitia, P. Monget, Genes encoding teleost Orthologs of human haploinsufficient and monoallelically expressed genes remain in duplicate more frequently than the whole genome, *Int. J. Genomics* 2021 (2021), 9028667, <https://doi.org/10.1155/2021/9028667>.
- [6] J.S. Taylor, J. Raes, Duplication and divergence: the evolution of new genes and old ideas, *Annu. Rev. Genet.* 38 (2004) 615–643, <https://doi.org/10.1146/annurev.genet.38.072902.092831>.
- [7] J. Nelson, T. Grande, M. Wilson, Fishes of the World, fifth ed., 2016, <https://doi.org/10.1002/9781119174844>.
- [8] I. Braasch, J.H. Postlethwait, Polyploidy in fish and the teleost genome duplication, in: P.S. Soltis, D.E. Soltis (Eds.), *Polyploidy and Genome Evolution*, Springer, Berlin, Heidelberg, 2012, pp. 341–383, [https://doi.org/10.1007/978-3-642-31442-1\\_17](https://doi.org/10.1007/978-3-642-31442-1_17).
- [9] S. Lien, et al., The Atlantic salmon genome provides insights into rediploidization, *Nature* 533 (7602) (2016) 200–205, <https://doi.org/10.1038/nature17164>, mai.
- [10] E. Parey, A. Louis, J. Montfort, Y. Guiguen, H.R. Crollius, C. Berthelot, An atlas of fish genome evolution reveals delayed rediploidization following the teleost whole-genome duplication, *Genome Res.* 32 (9) (2022) 1685–1697, <https://doi.org/10.1101/gr.276953.122>, sept.
- [11] D.J. Montefusco, N. Matmati, Y.A. Hannun, The yeast sphingolipid signaling landscape, *Chem. Phys. Lipids* 177 (2014) 26–40, <https://doi.org/10.1016/j.chph.2013.10.006>, janv.
- [12] H. Saito, Regulation of cross-talk in yeast MAPK signaling pathways, *Curr. Opin. Microbiol.* 13 (6) (2010), <https://doi.org/10.1016/j.mib.2010.09.001> déc.
- [13] P. Portela, S. Rossi, cAMP-PKA signal transduction specificity in *Saccharomyces cerevisiae*, *Curr. Genet.* 66 (6) (2020) 1093–1099, <https://doi.org/10.1007/s00294-020-01107-6>, déc.
- [14] A. Grandchamp, B. Piégut, P. Monget, Genes encoding teleost fish ligands and associated receptors remained in duplicate more frequently than the rest of the genome, *Genome Biol. Evol., avr* (2019), <https://doi.org/10.1093/gbe/evz078>.
- [15] R. Peto, Quantitative implications of the approximate irrelevance of mammalian body size and lifespan to lifelong cancer risk, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 370 (1673) (2015), 20150198, <https://doi.org/10.1098/rstb.2015.0198> juill.
- [16] M. Sulak, et al., TP53 copy number expansion is associated with the evolution of increased body size and an enhanced DNA damage response in elephants, *Elife* 5 (2016), e11994, <https://doi.org/10.7554/elife.11994> sept.
- [17] G.D. Bader, M.P. Cary, C. Sander, Pathguide: a pathway resource list, no Database issue, *Nucleic Acids Res.* 34 (2006) D504, <https://doi.org/10.1093/nar/gkj126>, 506, janv.
- [18] F. Cunningham, et al., Ensembl 2022, *Nucleic Acids Res.* 50 (D1) (2022) D988–D995, <https://doi.org/10.1093/nar/gkab1049>, janv.
- [19] C. Meslin, F. Brimau, P. Nagnan-Le Meillour, I. Callebaut, G. Pascal, P. Monget, The evolutionary history of the SAL1 gene family in eutherian mammals, *BMC Evol. Biol.* 11 (2011) 148, <https://doi.org/10.1186/1471-2148-11-148>, mai.
- [20] C. Meslin, et al., Evolution of genes involved in gamete interaction: evidence for positive selection, duplications and losses in vertebrates, *PLoS One* 7 (9) (2012), e44548, <https://doi.org/10.1371/journal.pone.0044548>.
- [21] C. Moros-Nicolás, S. Fouchécourt, G. Goudet, P. Monget, Genes encoding mammalian oviductal proteins involved in fertilization are subjected to gene death and positive selection, *J. Mol. Evol.* 86 (9) (2018) 655–667, <https://doi.org/10.1007/s00239-018-9878-0>, déc.
- [22] O. Jaillon, et al., Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype, *Nature* 431 (7011) (2004) 946–957, <https://doi.org/10.1038/nature03025>, oct.
- [23] T. Kon, et al., Single-cell transcriptomics of the goldfish retina reveals genetic divergence in the asymmetrically evolved subgenomes after allotetraploidization, *Commun. Biol.* 5 (1) (2022) 1404, <https://doi.org/10.1038/s42003-022-04351-3>, déc.
- [24] P. Xu, et al., The allotetraploid origin and asymmetrical genome evolution of the common carp *Cyprinus carpio*, *Nat. Commun.* 10 (1) (2019) 4625, <https://doi.org/10.1038/s41467-019-12644-1>, oct.

# Discussion

## 4.1 Étude sur les moments d'apparition

Dans un premier temps, notre étude a permis de mettre en lumière trois scénarios possibles sur la construction d'une voie de signalisation intracellulaire humaine. Notre premier scénario était que les voies sont arrivées dans le sens amont vers aval, donc des ligand-récepteur aux facteurs de transcription. Cette hypothèse nous semblait être la moins probable au vu des résultats de l'étude d'Anna Grandchamp durant sa thèse (**Grandchamp & Monget ; 2018**). En effet, ce travail a montré que 38% des récepteurs sont apparus avant leur ligand contre 21%, apparus après leur ligand. Or dans notre cas, le ligand est à la position 1 dans la cascade de signalisation, et le récepteur à la position 2 ce qui favoriserait la deuxième hypothèse. Cependant, nous avons travaillé sur un plus grand nombre d'espèces : 315 au lieu des 145 espèces dans l'étude précédente, et sur 25 clades au lieu de 10. Cette augmentation des données permet probablement d'être plus précis dans la recherche d'orthologues dans l'arbre de la vie, et d'être plus résolutifs concernant les noeuds d'apparition. Ce scénario pourrait expliquer 10 des voies KEGG.

Notre deuxième scénario concerne les voies qui se sont construites de l'aval vers l'amont, donc des facteurs de transcription vers les ligand-récepteur et a été validé pour 25 voies de notre étude ( $r < 0$ ,  $p$  value  $< 0.02$ ) dont 3 voies (Adipocytokine, Fc epsilon RI et VEGF) pour lesquelles la corrélation est particulièrement forte ( $r > 0.5$ ,  $p$  value  $< 3.15E-8$ ). Ce scénario peut également être motivé par d'autres éléments, comme le fait que les protéines en aval de voie de signalisation sont connectées avec plus de partenaires que les protéines en amont de la voie. La "multi-connectivité" d'une protéine augmente ses contraintes évolutives car cela imposerait une co-évolution à ses différents partenaires (**fraser \_ evolutionary \_ 2002 ; hahn \_ molecular \_ 2004 ; krylov \_ gene \_ 2003**).

Pour autant, cela pourrait également venir des types de fonction en amont des voies et en aval qui peuvent être différents. En effet, dans l'étude d'**alvarez-ponce \_ relationship \_ 2012**, il est montré une différence notable entre la nature des protéines en début et en fin de voie. Les protéines en amont d'une voie sont logiquement enrichies en activité kinases ( $q = 2,24 \times 10^{-4}$ ,  $q$  étant la  $p$ -value corrigée par Benjamini et Hochberg) tandis que les

protéines en aval sont enrichies en activités de transporteur transmembranaire ionique ( $q = 2,65 \times 10^{-5}$ ). De notre côté, nous n'avons pas regardé la *Gene Ontology* de nos protéines en début et en fin de voie, mais cela pourrait être reconsidéré.

Et notre dernier scénario concernait une absence d'ordre pour la construction d'une voie de signalisation. Ce scénario semble concerner 12 voies de notre étude.

Par ailleurs, l'étude d'Alvarez-Ponce qui comprenait 1049 protéines et 2436 interactions, a montré qu'il n'y avait pas d'incidence entre la position hiérarchique d'une protéine dans une voie et leur taux d'évolution. La particularité de l'étude était qu'il partait sans a priori de la voie c'est-à-dire qu'ils ont utilisés uniquement des interactions et non pas des voies définies comme le fait KEGG et la position hiérarchique était déterminée en fonction de leur position dans la voie (*upstream* ou *downstream*). Le taux d'évolution représente le ratio entre les mutations silencieuses et les mutations faux sens pour des séquences hommes vs. souris (**alvarez-ponce\_relationship\_2012**). Dans notre cas, le taux d'évolution (dN/dS) n'a pas été calculé. Nous pourrions également envisager de le faire pour notre étude.

Le taux d'évolution est donc différent du moment d'apparition, mais peut lui être lié. Plus un gène a un taux d'évolution élevé, par exemple s'il est soumis à une sélection positive, et plus il y a de chance qu'il ait beaucoup divergé jusqu'à devenir une sorte de nouveau gène et donc de ne pas être détecté en tant qu'orthologue. Dans ce cas, le moment d'apparition du gène humain pourrait d'être plus éloigné que pour un gène présentant une vitesse d'évolution plus lente.

Nous avons fait le choix durant cette étude de partir de l'homme, qui est l'espèce la plus référencée, et de partir d'un échantillon de voies de signalisation décrite comme telle sur la base de données KEGG.

De plus, nous avons imaginé 3 scénarios possibles, mais il existe d'autres scénarios auxquels nous n'avons exploré. On peut évoquer maintenant que des voies pourraient être construites à partir d'une chaîne d'interactions protéiques ancestrales « simple » et se serait vu complexifiée à travers l'évolution et les différentes spécificités des espèces. Cette réflexion s'est notamment faite en analysant de plus près la sous-voie RTK/RAS/ERK qui est présente dans 22 voies sur 47 de notre étude. Les éléments de cette sous-voie sont apparus tôt dans l'arbre de la vie puisque la voie est commune aux drosophiles, aux vers et aux humains (**ashton-beaucage\_signalisation\_2010**).

Cette première étude a également mis en lumière deux nœuds qui semblent être des pivots de l'évolution pour les gènes de voie de signalisation : le nœud des Opistoconte, et celui des Vertébrés. Ce sont ces mêmes nœuds qu'Anna Grandchamp avait mis en évidence dans son étude sur les ligand-récepteur (**grandchamp\_synchronous\_2018**). Le nœud des Opistoconte est un nœud peu surprenant au vu de la littérature. On constate que la comparaison levures-mammifères a montré de grandes similitudes de structure de réseaux moléculaires (Cross et al., 2011). Concernant le deuxième nœud évolutif notable, le nœud

des Vertébrés, il vient prendre son origine après deux duplications de génome, et il a été montré également que les gènes issus de duplication de génome avaient une évolution lente comparée aux gènes dupliqués à petite échelle (**satake\_evolution\_2012**). Dans cette même étude également, sont pointés du doigt les gènes d'expression ubiquiste qui évolueraient moins vite que les gènes d'expression tissulaires spécifiques. Et c'est effectivement un point qui pose une autre question : Est-ce que les gènes impliqués dans des voies de signalisation activées dans la majorité des cellules évoluent moins vites que des voies spécifiques de tissus ? Effectivement, certaines voies sont essentielles à toutes les cellules, notamment les voies menant à la croissance cellulaire ou la division cellulaire comme la voie cAMP (**sassone-corsi\_cyclic\_2012**) et d'autres spécifiques à une fonction comme les voies de l'immunité ou la voie Hippo régulatrice de la contraction musculaire entre autres grâce à sa capacité régulatrice de la taille des cellules (**zhao\_ippo\_2011**).

Un début de réponse peut être donné concernant les gènes des voies de l'immunité puisque nous avons trouvé que 72% des gènes des voies de l'immunité sont apparus après le noeud des Vertébrés. Ce constat rejoint par ailleurs la littérature qui évoque l'évolution rapide de ces gènes (**cooper\_evolution\_2006**; **schlesinger\_coevolutionary\_2014**).

Il aurait été intéressant de confronter différentes voies de signalisation présentes chez des espèces différentes de l'arbre de la vie des Opisthocontes. Seulement, la voie MAPK est l'unique voie est disponible chez la levure sur KEGG.

Il faut tout de même émettre des réserves sur nos résultats car un moment d'apparition pour une interaction ne signifie pas qu'elle est fonctionnelle. Tout comme un moment d'apparition d'un gène ne rend pas la protéine fonctionnelle chez toutes les espèces concernées. Il faudrait, pour confirmer ces résultats, valider fonctionnellement les interactions chez les espèces avec l'ancêtre commun le plus éloigné de l'homme par un système de double hybride par exemple. Un système de double hybride permettrait de valider une interaction entre deux protéines, mais dans des conditions artificielles, pouvant être très éloignée de la réalité. Ce qui serait un travail colossal et très coûteux pour le nombre d'interactions et le nombre d'espèces différentes. On pourrait également utiliser des techniques novatrices et relativement inexplorées comme la résurrection des gènes. Des travaux ont été entamés sur la résurrection de protéines afin de mieux connaître les fonctions des gènes ancestraux. La technique consiste à réintégrer dans un organisme un gène jusqu'alors inactif ou désactivé chez cette espèce par le biais d'un système tel que Crispr-Cas9. Bien que ce soit une technique complexe, elle se veut prometteuse (**harms\_evolutionary\_2013**). Une autre façon de répondre à ces questions serait de modéliser les structures tridimensionnelles des protéines ancestrales. La technologie AlphaFold2 est une avancée majeure dans le domaine de la prédiction de la structure des protéines (**cramer\_alphafold2\_2021**). Avec cet outil, nous pourrions pousser l'étude d'un point de vue tridimensionnel et se demander pour chacune des protéines des voies de signalisation, et à travers l'arbre de la vie, à quel moment ont eu lieu les modifications tridimensionnelles, si changement il y a. Et

pourquoi ne pas rêver d'arbres phylogénétiques de gènes basés non pas sur des homologies de séquences primaires, mais sur des homologies de structure 3D des protéines ?

Les gènes et leur évolution sont des sujets qui animent la communauté scientifique. Par ailleurs, l'évolution est constante et perpétuelle, ce qui mène à des questionnements peut être plus hardis. Par exemple, on pourrait se poser la question de futures duplications de génome chez certains clades. Certaines espèces vivant dans des régions arides ont peut-être déjà doublé certains de leurs gènes pour mieux supporter la chaleur. Et si les températures continuent d'augmenter, il est possible que, sur plusieurs siècles, les poissons développent des duplications génétiques pour accroître leur résistance à la chaleur. De même, ils pourraient évoluer pour mieux s'adapter à des environnements plus salins si les rivières venaient à s'assécher complètement. Grâce au système de Crispr-Cas9, nous arrivons de mieux en mieux à modifier les génomes précisément. Il pourrait donc être envisageable de dupliquer des génomes. Mais ceci reste encore très utopique.

Dans un registre beaucoup plus réalisable, il serait intéressant de regarder plus en profondeur les gènes orphelins, car l'étude a montré qu'une majorité protéines en interactions attendaient leur partenaire (84% d'interactions asynchrones). Que font les protéines sans leur partenaire ? Premièrement, nous avons choisi de centrer notre étude sur les gènes des voies de signalisation qui représentent 3 000 interactions protéine-protéine unique, ce qui en fait un petit échantillon sur les 93 000 interactions protéine-protéine uniques recensées par **luck\_proteome-scale\_2017**. Et deuxièmement, des études ont été menées sur des récepteurs nucléaires orphelins, mais les résultats sont trop divergents, ce qui montre bien la difficulté de la problématique (**markov\_origin\_2011**).

## 4.2 Étude sur les téléostéens

Dans un deuxième temps, nous avons étudié la stœchiométrie des interactions chez les espèces 3 WGD et 4 WGD du clade des téléostéens. Cette étude a montré que pour l'ensemble des espèces de poissons téléostéens, la stœchiométrie des interactions était respectée en fonction du nombre de duplications de génome des espèces. Pour autant, une notion n'a pas été prise en compte : les paralogues retrouvés sont-ils des paralogues ohnologues, ou des paralogues « simples ». On recense environ 26% de gènes ohnologues chez le poisson zèbre ([howe\\_zebrafish\\_2013](#)) et environ 44% de gènes ohnologues chez les salmonidés ([dimos\\_homology\\_2023](#)).

Ce que nous avons trouvé étonnant, ce sont les résultats concernant les espèces 4 WGD. Pour rappel, les clades des salmonidés et des carpes ont vécu indépendamment une 4ème duplication de génome, à des temps différents et pour autant, les résultats concernant les différentes stœchiométries possibles pour ces deux groupes d'espèces étaient sensiblement les mêmes. La 4ème duplication de génome des Salmonidés a eu lieu, il y a environ 80 millions d'années ([lien\\_atlantic\\_2016](#)), tandis que celle des carpes a eu lieu il y a environ 14 millions d'années ([jaillon\\_genome\\_2004](#); [kon\\_single-cell\\_2022](#); [xu\\_allotetraploid\\_2019](#)). Nous avons constaté des proportions similaires concernant les différentes stœchiométries, mais nous n'avons pas étudié si les interactions en question étaient les mêmes pour les deux clades indépendants. Nous pourrions lister ces interactions afin de mieux comprendre si la pression de maintien des interactions en stœchiométrie respectées étaient les mêmes malgré une divergence aussi lointaine.

De plus, il serait intéressant d'avoir une approche par voie pour les téléostéens, c'est-à-dire regarder si des sous-voies au sein des voies sont en stœchiométrie respectée. Dans notre cas, nous avons étudié interaction par interaction et non directement en utilisant la voie. C'était dû à un manque de maîtrise des outils igraph au moment de l'étude, mais ça pourrait mettre en lumière une pression de maintien en duplicit ou même de retour en singleton par la voie au sein des téléostéens. Ça pourrait donner également des indications sur les voies de signalisation existantes chez les téléostéens qui sont encore assez peu étudiées.

Nous avons également noté une part non négligeable de gènes absents chez les téléostéens. Les interactions partiellement ou totalement perdues chez les téléostéens étaient en moyenne à hauteur de 25%. Dans quelques cas, notamment dans le cas où nous avions une perte chez quelques espèces tandis que l'interaction était présente chez le reste des espèces, nous avons retrouvé des traces de pseudogènes. Mais, dans le cas où nous ne retrouvions pas l'interaction pour l'ensemble du clade, nous avons émis l'hypothèse que les gènes et donc les interactions seraient nées après. Seulement, il serait pertinent de croiser nos deux études et de regarder si les gènes étaient apparus avant le clade des Vertébrés, mais se seraient éteints chez les téléostéens car non essentiel aux espèces. De plus, cette

hypothèse est un biais de notre étude que nous pourrions améliorer en redéfinissant les conditions d'admission d'un noeud d'apparition. Nous pourrions convenir qu'il faut plus de 80% (par exemple) de présence du gène dans l'ensemble des espèces entre l'homme et le noeud le plus éloigné où on retrouve un orthologue.

## 4.3 Limite du matériel et de la méthode

Durant ces différentes études, nous avons fait des choix de matériel à utiliser et d'une méthodologie à suivre. Nous avons longuement hésité entre une étude exhaustive avec toutes les voies de signalisation que proposait KEGG, et même pousser l'exhaustivité et regrouper des données de différentes bases de données comme celles de la base de données BioGrid, String, Reactome ou encore NetPath qui sont également des bases de données d'interactions protéiques, ou faire une étude plus spécifique sur une voie de signalisation précise qui pourrait notamment intéresser l'équipe de recherche dans laquelle j'ai réalisé ma thèse. Nous avons fini par trancher sur une liste de 47 voies de signalisation sur la base de données KEGG, car c'était la mieux référencée (**rigden\_26th\_2019**).

Concernant l'arbre et les espèces utilisées, Anna Grandchamp avait utilisé ces deux mêmes bases de données. Cette thèse s'inscrivant naturellement à la suite de celle d'Anna Grandchamp en 2018, il y avait un souhait de confronter nos résultats aux siens et donc d'utiliser les mêmes outils. Pour autant, la base de données Ensembl et par extension Genomicus a connu un changement important entre la version 93 (janvier 2018) et la version 94 (juillet 2018). En version 93, les arbres étaient présentés de telle façon à ce que l'ensemble des paralogues et orthologues d'un gène soient présents dans un arbre unique. Seulement, avec les ajouts de nouvelles espèces séquencées, la base de données a changé de méthodes de clustering des gènes et de nombreux liens de paralogie ont disparu, rendant impossible la visualisation des grandes familles de gènes. Les arbres ont donc été redécoupés en sous-arbre (**emily\_changes\_2018**). De ces changements majeurs, nous nous sommes longuement posé la question de savoir si nous devions rester sur les arbres de version 93, ou si nous prenions les versions les plus récentes. Pour nos questions de recherche, nous avons décidé de prendre les plus récentes versions afin d'avoir le plus d'espèces possibles et de recouvrir au maximum l'arbre de la vie des Opisthocontes et plus précisément des téléostéens pour l'étude 2.

Et afin de valider un moment d'apparition, il pourrait être réalisable de faire des BLAST pour chacun des gènes avec l'espèce ancestrale la plus éloignée d'un point de vue évolutif.

De plus, nous avons été confrontés à de nombreux questionnements concernant la gestion des données KEGG. En effet, KEGG reste une base de données renseignée à la main, et par facilité de compréhension, des voies peuvent afficher une famille de gènes paralogues derrière une étiquette portant un nom plus générique. Seulement, elle ne le fait pas systématiquement et des familles de gènes peuvent se retrouver dans des étiquettes uniques, comme c'est le cas par exemple pour la voie PPAR, les gènes PPAR alpha, beta/delta et gamma sont des paralogues qui sont affichés distinctement dans la voie. L'hypothèse émise était que les paralogues ont des fonctions différentes et des interactions spécifiques comme des interactions communes entre elles, comme semble le montrer la voie.

Pour autant, il a fallu décider de ce que l'on faisait dans des cas comme PPAR. Nous nous sommes également demandé s'il n'était pas préférable de gérer tous les paralogues de la même façon et de prendre le même ancêtre commun pour une même famille de gènes. Là encore, nous avons discuté pour savoir quel ancêtre commun nous prenions si les arbres des différents paralogues n'étaient pas d'accord. Prenions-nous le plus ancien, le plus récent, le plus représenté ? Nous avons fini par faire l'étude avec toutes les configurations possibles avant de trancher sur celle qui nous semblait présenter le moins de biais : la gestion des paralogues telle que présenté par KEGG, soit en groupe, soit unique. Et pour les gènes avec de nombreux paralogues en groupe, de prendre le nœud ancestral le plus éloigné.

Par souci de simplification des voies, nous avons également omis l'information de l'interaction, à savoir si l'interaction était activatrice, ou inhibitrice, ou autre. Cependant, il faudrait confronter nos résultats à cette dimension-là. Nous pourrions séparer les interactions en fonction de leur type et regarder si des corrélations se dégagent.

## 4.4 Perspectives

Les résultats que nous avons obtenus concernant le moment d'apparition des gènes peuvent faire l'objet de projets complémentaires. J'aimerais ajuster plus précisément les moments d'apparition en regardant chez les clades plus éloignés encore comme celui des plantes par exemple ou celui des bactéries. Le génome de l'homme comportant environ 25% de ses gènes en commun avec les plantes, il est fort possible qu'un certain nombre de gènes soient des gènes de voie de signalisation.

Par ailleurs, nous aimeraisons proposer à la base de données KEGG d'implémenter un nouvel outil faisant apparaître les moments d'apparitions des gènes directement sur les voies avec un système de couleur comme nous l'avons réalisé. En poussant nos idées, nous pourrions également afficher le nombre d'orthologues retrouvés pour une espèce donnée, cela permettrait en un coup d'œil de connaître la présence ou non d'un gène chez une espèce, et de connaître la stoechiométrie de ces interactions dans un cadre de voie de signalisation.

Nos résultats concernant la coévolution de gènes impliqués en interactions nous ont donné envie de regarder ce qu'il en était concernant les gènes pro-apoptotiques et anti-apoptotiques. En effet, il serait intéressant de regarder qui de l'un ou de l'autre apparaît le premier. Cela rejoindrait les études précédentes sur les ligand-récepteur, et les gènes des voies de signalisation.

Bien évidemment, il serait plus juste scientifiquement de valider tous nos résultats avec des approches de biologie expérimentale. Cependant je n'ai pas les outils pour imaginer des expériences poussées, et de plus, en travaillant sur un si large matériel comme nous l'avons fait ici, il faudrait faire des choix.

# Conclusion

Nos travaux ont testé trois scénarios possibles de construction des voies de signalisation. Nous mettons également en lumière deux moments d'apparition clés dans l'arbre de la vie pour les gènes impliqués dans une voie de signalisation qui sont le nœud des Opisthoconte et celui des Vertébrés. Au regard de ces résultats, la double duplication de génomes à l'émergence des Vertébrés peut être à l'origine de la naissance de nouveaux gènes de voie de signalisation. Nous avons poussé nos recherches en détail pour connaître le comportement des interactions chez des clades les plus anciens des vertébrés, à savoir les téléostéens qui ont subi 3 ou 4 duplications de génome. Nous avons constaté que les orthologues téléostéens des gènes humains des voies de signalisation restaient plus souvent en duplicitat ou en triplicat chez les espèces 3 ou 4 WGD téléostéens que les gènes du génome. Et nous avons retrouvé une majorité d'interactions protéine-protéine à stoechiométrie respectée en moyenne chez les téléostéens.



---

## Annexes

## 6.1 Article 3

## Research Article

# Genes Encoding Teleost Orthologs of Human Haploinsufficient and Monoallelically Expressed Genes Remain in Duplicate More Frequently Than the Whole Genome

Floriane Picolo,<sup>1</sup> Anna Grandchamp,<sup>1</sup> Benoît Piégu,<sup>1</sup> Antoine D. Rolland<sup>2</sup>,  
Reiner A. Veitia,<sup>3,4,5</sup> and Philippe Monget<sup>1</sup>

<sup>1</sup>PRC, UMR85, INRAE, CNRS, IFCE, Université de Tours, F-37380 Nouzilly, France

<sup>2</sup>Univ Rennes, Inserm, EHESP, Iset (Institut de Recherche en Santé, Environnement et Travail)-UMR S 1085, F-35000 Rennes, France

<sup>3</sup>Université de Paris, F-75006 Paris, France

<sup>4</sup>Université de Paris, CNRS, Institut Jacques Monod, F-75006 Paris, France

<sup>5</sup>Université Paris-Saclay, Institut de Biologie F. Jacob, Commissariat à l'Energie Atomique, Fontenay aux Roses, France

Correspondence should be addressed to Philippe Monget; philippe.monget@inrae.fr

Received 14 April 2021; Accepted 5 July 2021; Published 30 July 2021

Academic Editor: Mohamed Salem

Copyright © 2021 Floriane Picolo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Gene dosage is an important issue both in cell and evolutionary biology. Most genes are present in two copies or alleles in diploid eukaryotic cells. The most outstanding exception is monoallelic gene expression (MA) that concerns genes localized on the X chromosome or in regions undergoing parental imprinting in eutherians, and many other genes scattered throughout the genome. In diploids, haploinsufficiency (HI) implies that a single functional copy of a gene in a diploid organism is insufficient to ensure a normal biological function. One of the most important mechanisms ensuring functional innovation during evolution is whole genome duplication (WGD). In addition to the two WGDs that have occurred in vertebrate genomes, the teleost genomes underwent an additional WGD, after their divergence from tetrapods. In the present work, we have studied on 57 teleost species whether the orthologs of human MA or HI genes remain more frequently in duplicates or returned more frequently in singleton than the rest of the genome. Our results show that the teleost orthologs of HI human genes remained more frequently in duplicate than the rest of the genome in all of the teleost species studied. No signal was observed for the orthologs of genes mapping to the human X chromosome or subjected to parental imprinting. Surprisingly, the teleost orthologs of the other human MA genes remained in duplicate more frequently than the rest of the genome for most teleost species. These results suggest that the teleost orthologs of MA and HI human genes also undergo selective pressures either related to absolute protein amounts and/or of dosage balance issues. However, these constraints seem to be different for MA genes in teleost in comparison with human genomes.

## 1. Introduction

Gene dosage effects are an important phenomenon in cell biology that has evolutionary consequences. Indeed, in diploid eukaryotic cells, most genes are present in two copies that are transcribed and produce functional proteins. However, there are exceptions. The most outstanding exception is the case of monoallelic gene expression (MA). This is so for the majority of genes that are present on the X chromosome of eutherian

mammals, genes that present a parental imprinting in eutherians, and genes encoding immunoglobulins and olfactory receptors [1]. Monoallelic expression of genes is under an epigenetic control that is not well understood. For these genes, dysregulation of the mechanism(s) underlying monoallelic expression can lead to expression of both alleles and to overexpression of the corresponding protein and thus to severe pathologies [2]. An abnormal situation concerns haploinsufficiency. Haploinsufficiency is a biological phenomenon

responsible for the fact that a single functional copy of a gene in a diploid organism is insufficient to ensure a normal biological function. Haploinsufficiency is detected more frequently in essential genes than in nonessential genes in yeast [3]. Two nonmutually exclusive theories have been proposed to explain the cause of haploinsufficiency: the “insufficient amounts” hypothesis and the gene dosage balance hypothesis (GDBH). The “insufficient amounts” hypothesis states that haploinsufficiency is the consequence of a reduced protein amount due to the loss of function of one allele, this amount being insufficient to ensure its biological function [4]. This hypothesis does not explain why haploinsufficiency persisted over evolutionary time. The GDBH suggests that the phenotype caused by changes of protein level in a biological process is due to stoichiometric imbalances in protein complexes or cellular circuits involved in cellular functions [5, 6]. This hypothesis predicts that haploinsufficient genes are responsible for a biological defect when the amount of proteins is halved (such as A in a complex A-B-A) but also in excess in particular cases (such as B in the same complex) [6]. In contrast to the “insufficient amounts” hypothesis, this hypothesis proposes an explanation for the conservation of haploinsufficiency during evolution.

One of the most important mechanisms ensuring functional innovation during evolution is gene duplication or the duplication of entire genome [7, 8]. Whole genome duplication (WGD) events have been observed in all taxonomic groups: bacteria [9], unicellular eukaryotes [10], and plants [11]. In vertebrates, there have been two rounds of duplication of the ancestral deuterostome genome [12]. One of the striking features that characterize the teleost genomes is that they underwent an additional WGD, also called the teleost-specific genome duplication (TGD), after the divergence from tetrapod [13]. This specific WGD event provided important additional genetic material, which strongly contributed to the radiation of teleost fishes [14]. Teleost constitutes a monophyletic group of ray finned fishes and is the largest and most diverse group of vertebrates [15–18]. The high diversity of fish species combined with a recent complete duplication makes Clupeocephala a group of great interest for the study of complete genome duplication in the animal kingdom.

Unlike single-gene duplication events, a WGD provides all at once a large number of new genetic material, promoting an increased inter- and intraspecific diversity [19, 20]. Interestingly, after WGD, all genes do not remain in duplicate with the same probability. Most models predict a rapid return of part of the duplicates to a singleton state [21], the extra-copies being rapidly pseudogenized [22]. In particular for the rainbow trout, whose genome has duplicated one more time than that of the teleost about 100 my ago, it is estimated that about 48% of the genome remained in duplicate, when the remaining 52% of the genome quickly returned to a singleton state [23].

Understanding the rules explaining why certain genes remain in duplicate when others return to singleton is a challenging issue. It has been shown that certain families of genes are more likely to remain as duplicates in all taxonomic groups studied. This is the case for transcription factors, pro-

tein kinases, enzymes, and transporters [24]. Recently, we showed that this is also the case for genes encoding membrane receptors and their ligands [25]. The first explanation that has been put forward to explain the fact that genes are more often kept in duplicate is that these molecules are involved in key functions common to all organisms. Their quantitative increase would favor these key functions because of an increase in the number of molecules produced (selection for an absolute dosage increase) and/or because of a compensation of a potential loss of function mutation of one of both copies. Another explanation is based on the respect of gene dosage balance. This is particularly so for proteins whose function is heavily dependent on interactions with partners.

In the present work, we have studied on 57 teleost species whether the orthologs of human genes known to present a monoallelic (MA) expression or to be haploinsufficient (HI) in human remain more frequently in duplicate or returned more frequently to as singleton state than the whole genome in fish species or not.

## 2. Results and Discussion

We found a mean number of 13882 human genes on 22836 (60.8%) that possess at least one ortholog in at least one teleost genome. Among them, an average of 9854 (ranging from 3530 to 10868) have returned in singleton, an average of 3135 (ranging from 2323 to 7066) remained in duplicate, and an average of 893 (ranging from 337 to 4772) are in triplicate or more copies.

Concerning the 312 human HI genes, 299 (95.8%) possessed at least one ortholog in at least one teleost genome. Among them, an average of 172 (ranging from 47 to 199 depending on the studied species) have returned to singleton, an average of 85 (ranging from 68 to 122 depending on the species) remained in duplicate, and an average of 19 (ranging from 3 to 140) are in triplicate or more copies. A total of 285 genes remained in duplicate (or more) in at least one species among the 57 teleost species studied (Figure 1 and Suppl. Data (available here)). In comparison with the whole genome, this higher percentage of genes returned to singleton and remained in duplicate or more is significantly different for 55 species out of 57 (chi<sup>2</sup> analysis, *p* value ranging from 0.058 to 4.2E – 6) and for the 57 species studied (according to a hypergeometric test, *p* value ranging from 0.034 to 8.5E – 6; Suppl. data). Moreover, in comparison with the whole genome as well, the higher percentage of genes that are in triplicate or more copies is significantly higher in the genomes of rainbow trout, brown trout, Atlantic salmon, huchen, and common carp (*p* value ranging from 1.3E – 8 to 8.1E – 4) but not in the genome of the other teleosts. These results suggest that the teleost orthologs of HI human genes are also subjected to selective pressures either related to absolute protein amounts and/or of dosage balance issues. This suggests that HI genes in humans undergo similar constraints in teleost.

Among the 285 genes that remained in duplicate in at least one teleost species, 76 genes remained in duplicate or more in at least 80% (45) of the species. These genes encode

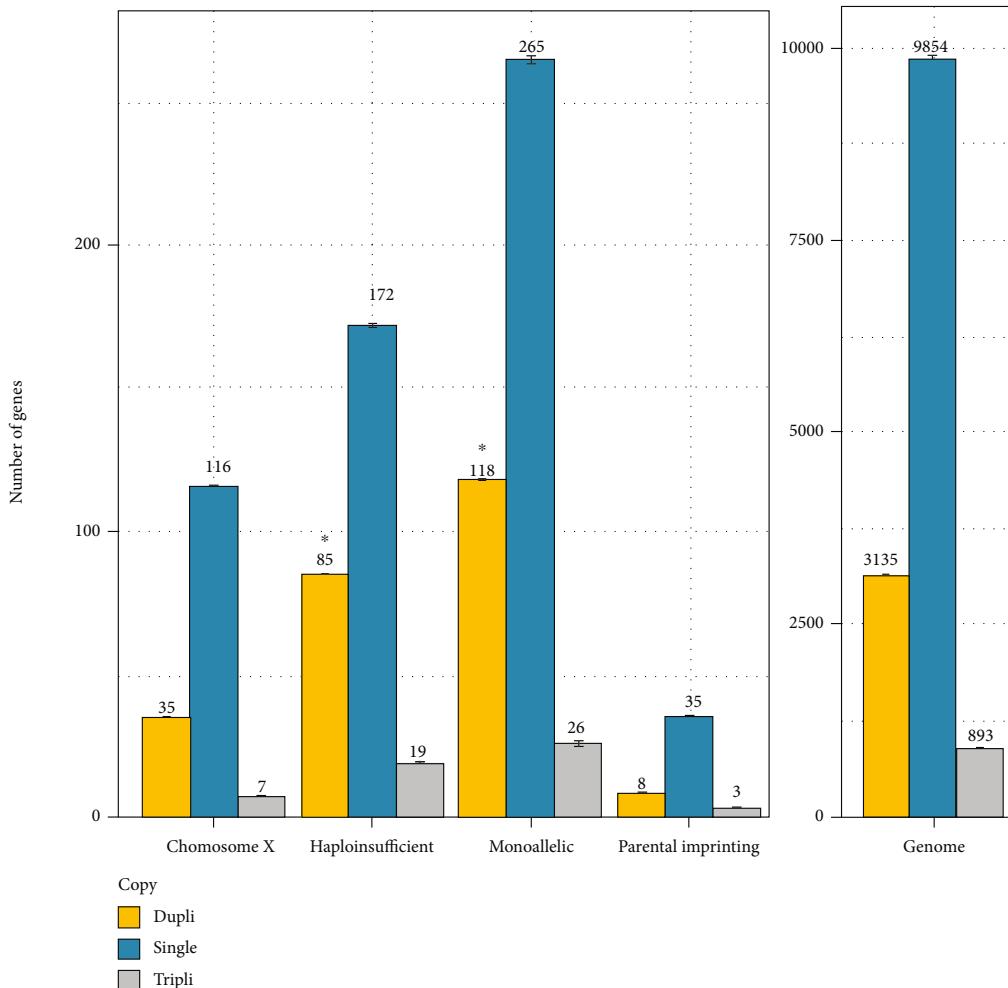


FIGURE 1: Barplot of the global distribution of the genes in each category: teleost orthologs of human genes mapping to the X chromosome, of human haploinsufficient (HI) genes, of human genes of monoallelic expression (MA, except genes that present a parental imprinting and localized on the X chromosome), and of human genes that present a parental imprinting. Right: teleost orthologs of human genes of the whole genome. The yellow bars correspond to the genes that remained in duplicate; the blue bars correspond to the genes returned to singleton. The grey bars correspond to the genes in triplicate or more. The results are presented as the mean  $\pm$  SEM. \* indicates a significant difference compared with the whole genome ( $p < 0.05$ ).

more (from 3 to 38 more times) transcription factors than the rest of the genome: bHLH transcription factor binding (Gene Ontology/GO:0043425); RNA polymerase II activating transcription factor binding (GO:0001102); activating transcription factor binding (GO:0033613); transcription factor binding (GO:0008134); DNA-binding transcription factor binding (GO:0140297); DNA-binding transcription factor activity and RNA polymerase II-specific (GO:0000981); and DNA-binding transcription factor activity (GO:0003700). This enrichment of GO terms is completely in accordance with previously reported findings [6]. There was no particular representative GO among the genes retained as triplicates in the genome of teleost species. These results are compatible

both with direct insufficiency of a transcription factor and with balance issues (as they are often multisubunit complexes). Threshold effects can also be at play because of the strongly nonlinear relationships (sigmoidal or S-shaped) produced by the cooperative binding of a transcription factor to a cis-regulatory sequence and the transcriptional response. Thus, depending on the concentration of transcription factor, a halved dosage may not be sufficient to cross the threshold required for a normal transcriptional response [6].

Concerning the 206 X-linked human genes, 176 (82.6%) possessed at least one ortholog in at least one teleost genome. Among them, an average of 116 (ranging from 32 to 132 depending on the studied species) have returned to singleton,

an average of 35 (ranging from 23 to 79 depending on the species) remained in duplicate, and an average of 7 (ranging from 0 to 54) are in triplicate or more copies (Figure 1 and Suppl. Data). Concerning the 90 imprinted genes, 51 (56.7%) had at least one ortholog in at least one teleost genome. Among them, an average of 35 (ranging from 12 to 41 depending on the studied species) have returned to singleton, an average of 8 (ranging from 3 to 23 depending on the species) remained in duplicate, and an average of 3 (ranging from 0 to 20) are in triplicate or more copies (Figure 1 and Suppl. Data). Thus, the teleost orthologs of human genes subjected to genetic imprinting or located on the human X chromosome returned to singleton or remained in duplicate (or remain present as triplicates or more copies), in the same proportions than the rest of the genome.

Concerning the 580 human MA genes that are not on the X chromosome and that are not subjected to parental imprinting, 469 (80.9%) had at least one ortholog in at least one teleost genome. Among them, an average of 265 (ranging from 87 to 296) have returned to singleton, an average of 118 (ranging from 87 to 193) remained in duplicate, and an average of 26 (ranging from 4 to 160) were found in triplicate or more copies. A total of 437 genes remained in duplicate in at least one species among the 57 teleost species studied (Figure 1 and Suppl. Data). In comparison with the whole genome, the difference of percentage of genes remained in duplicate or more is significantly higher for 47 species on 57 ( $\chi^2$  analysis,  $p$  value ranging from 0.055 to 6.5 4) and for 50 species on 57 (hypergeometric test,  $p$  value ranging from 0.044 to 6.2 4; Suppl. data). Moreover, in comparison with the whole genome as well, the difference of percentage of genes that are in triplicate or more copies is significantly higher in the genomes of rainbow trout, brown trout, Atlantic salmon, huchen, and common carp ( $p$  value ranging from 0.056 to 5.3 3), not in the genome of the other teleosts. Whether these results are generalizable to other salmonids needs further investigation. We found this result surprising. Indeed, one would have hypothesized that the teleost orthologs of MA human genes returned more frequently to singleton than the whole genome. This suggests that the regulation (epigenetic mechanism) of monoallelic expression is not likely to occur for these genes in teleosts. Moreover, this suggests that the constraints to express only one allele in the human do not exist for these genes in teleosts. That being said, MA is a complex regulatory process that has evolved perhaps due to parent-offspring conflict. It might be possible that genes that need such control in mammals would be those with a general need for dosage balance in other species like teleost. Unlike the HI genes, there was no particularly representative GO among the MA genes.

### 3. Material and Methods

We studied 57 species of fish: Amazon molly (*Poecilia formosa*), Atlantic herring (*Clupea harengus*), Atlantic salmon (*Salmo salar*), ballan wrasse (*Labrus bergylta*), barramundi perch (*Lates calcarifer*), blue tilapia (*Oreochromis aureus*), blunt-snouted clingfish (*Gouania willdenowi*), brown trout (*Salmo trutta*), Burton's mouthbrooder (*Haplochromis bur-*

*toni*), channel bull blenny (*Cottoperca gobio*), channel catfish (*Ictalurus punctatus*), climbing perch (*Anabas testudineus*), cod (*Gadus morhua*), common carp (*Cyprinus carpio common\_carp\_genome*), denticle herring (*Denticeps clupeoides*), Eastern happy (*Astatotilapia calliptera*), electric eel (*Electrophorus electricus*), European seabass (*Dicentrarchus labrax*), fugu (*Takifugu rubripes*), gilthead seabream (*Sparus aurata*), greater amberjack (*Seriola dumerili*), guppy (*Poecilia reticulata*), huchen (*Hucho hucho*), Indian glassy fish (*Parambassis ranga*), Indian medaka (*Oryzias melastigma*), Japanese medaka HdrR (*Oryzias latipes ASM223467v1*), Japanese medaka HNI (*Oryzias latipes ASM223471v1*), Japanese medaka HSOK (*Oryzias latipes ASM223469v1*), jewelled blenny (*Salarias fasciatus*), large yellow croaker (*Larimichthys crocea*), live sharksucker (*Echeneis naucrates*), lyretail cichlid (*Neolamprologus brichardi*), Makobe Island cichlid (*Pundamilia nyererei*), Mexican tetra (*Astyanax mexicanus Astyanax\_mexicanus-2.0*), Midas cichlid (*Amphilophus citrinellus*), mummichog (*Fundulus heteroclitus*), Nile tilapia (*Oreochromis niloticus*), northern pike (*Esox lucius*), orbicular cardinalfish (*Sphaeramia orbicularis*), Pachon cavefish (*Astyanax mexicanus Astyanax\_mexicanus-1.0.2*), pinecone soldierfish (*Myripristis murdjan*), rainbow trout (*Oncorhynchus mykiss*), red-bellied piranha (*Pygocentrus nattereri*), sailfin molly (*Poecilia latipinna*), sheephead minnow (*Cyprinodon variegatus*), shortfin molly (*Poecilia mexicana*), siamese fighting fish (*Betta splendens*), stickleback (*Gasterosteus aculeatus*), swamp eel (*Monopterus albus*), tetraodon (*Tetraodon nigroviridis*), tiger tail seahorse (*Hippocampus comes*), tongue sole (*Cynoglossus semilaevis*), turbot (*Scophthalmus maximus*), yellowtail amberjack (*Seriola lalandi dorsalis*), zebra mbuna (*Maylandia zebra*), zebrafish (*Danio rerio*), and zig-zag eel (*Mastacembelus armatus*).

These fish species diverged after complete TGD. The human genes were retrieved from ENSEMBL. The ortholog copy for each gene was established in every one of the 57 fish species. Then, in each species, the fate (singleton vs duplicate) of the entirety of the human gene orthologs was studied. Moreover, a total of 312 human genes known to be haploinsufficient were recovered from Clingene (<https://www.ncbi.nlm.nih.gov/projects/dbvar/clingen/>), 580 human genes were known to be monoallelic [26], 206 X human chromosome genes were recovered for GeneImprint (<http://www.geneimprint.com/site/genes-by-species>), and there are 90 genetic imprinting genes [27], and the fate of their fish orthologs was recovered. A list of human genes (GRCh38.p13) was generated using BioMart from Ensembl Genes 101. The set of human genes encoding a protein (protein\_coding) is selected from the gene type filter. The selected attributes in the homologous category are the different species of teleostans listed in ENSEMBL. Only stable gene IDs were selected. A list of 22836 human genes encoding a protein is listed.

We got between 12,918 (tetraodon) and 14,626 (brown trout) orthologous genes by fish species (average: 13,882). This does not represent the entire genome of each fish but allowed us to make strong statistics. Moreover, we compared the global evolution of the whole human genome that had orthologs in fishes with the specific evolution of human

MA and HI genes in fish species. We studied whether these fish orthologs of MA and HI genes remained as a duplicate copy or had return to singleton in the same proportion as whole human ortholog genes.

Both the  $\chi^2$  test statistical analysis and hypergeometric analysis with Benjamini-Hochberg correction were used to test the hypothesis that teleost genes that are orthologs of human MA and HI genes remained more in duplicate than the whole genome. All the statistical tests conducted in our study were performed in R. Moreover, the Panther database (<http://www.pantherdb.org/>) was used to study the gene ontology of teleost genes that are orthologs to human HI genes, and Fisher's test with Benjamini-Hochberg correction was used to classify genes according to the family.

### Data Availability

The underlying data supporting the results of this study can essentially be found in Supplemental data and can be verified on ENSEMBL (<http://www.ensembl.org/index.html>).

### Disclosure

This manuscript has been submitted as a preprint at <https://www.biorxiv.org/content/10.1101/2021.01.12.426466v1>.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

### Acknowledgments

The present study was supported by a fellowship from the French Ministry of Research, by the INRAE institute, and by a thesis scholarship funded by the University of Tours (France). Thanks are also due to Alexandra Louis and Hugues Roest Crollius for helpful discussion and technical assistance.

### Supplementary Materials

Tables of statistic tests for each species of teleost and for each category. Each category is represented in each of the four Suppl. Data (HI, MA, X chromosome, and parental imprinting), and these four tables have the same construction. By column (category HI for example): (A) species; (B) total number of teleost genes returned in singleton; (C) total number of teleost genes remained in duplicate; (D) total number of teleost genes in triplicate or more copies; (E) total number of teleost genes with a human ortholog; (F) number of teleost orthologs of HI human genes returned in singleton; (G) number of teleost orthologs of HI human genes remained in duplicate; (H) number of teleost orthologs of HI human genes in triplicate or more copies; (I) total number of teleost orthologs to human HI gene; (J)  $\chi^2$  value of the repartition of HI orthologs in singleton, in duplicate or more copies in comparison with the whole genome; (K)  $p$  value of  $\chi^2$  test; (L)  $\chi^2$  false discovery rate (FDR) by Benjamini Hochberg (BH) procedure. (M)  $p$  value of hypergeometric test between singleton and duplicate/more copies; (N) hypergeometric

FDR by BH procedure; (O)  $p$  value of hypergeometric test between triplicate or more copies and duplicate or less copies; (P) hypergeometric FDR by BH procedure. The same organization of columns is used for the other categories (MA, X chromosome, and imprinted genes). Concerning for HI and MA categories, the  $\chi^2$  test is significant for 55/57 and 47/57 species, respectively, and the hypergeometric test is significant for 57/57 and 50/57 species, respectively; i.e., these orthologs remain more frequently in duplicate than the whole genome. For comparison between triplicate (or more copies) and duplicate (or less copies), the hypergeometric test is significant for 5/57 (salmonids and carp) and 3/57 species, respectively (salmonid and carp as well). (*Supplementary Materials*)

### References

- [1] A. Chess, "Monoallelic gene expression in mammals," *Annual Review of Genetics*, vol. 50, no. 1, pp. 317–327, 2016.
- [2] B. Horsthemke, "Mechanisms of imprint dysregulation," *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, vol. 154C, no. 3, pp. 321–328, 2010.
- [3] S. Ohnuki and Y. Ohya, "High-dimensional single-cell phenotyping reveals extensive haploinsufficiency," *PLoS Biology*, vol. 16, no. 5, article e2005130, 2018.
- [4] A. M. Deutschbauer, D. F. Jaramillo, M. Proctor et al., "Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast," *Genetics*, vol. 169, no. 4, pp. 1915–1925, 2005.
- [5] B. Papp, C. Pál, and L. D. Hurst, "Dosage sensitivity and the evolution of gene families in yeast," *Nature*, vol. 424, no. 6945, pp. 194–197, 2003.
- [6] R. A. Veitia, "Exploring the etiology of haploinsufficiency," *Bio Essays: News and Reviews in Molecular, Cellular and Developmental Biology*, vol. 24, no. 2, pp. 175–184, 2002.
- [7] H. Innan and F. Kondrashov, "The evolution of gene duplications: classifying and distinguishing between models," *Nature Reviews Genetics*, vol. 11, no. 2, pp. 97–108, 2010.
- [8] S. Ohno, U. Wolf, and N. B. Atkin, "Evolution from fish to mammals by gene duplication," *Hereditas*, vol. 59, no. 1, pp. 169–187, 1968.
- [9] M. Kuroda, T. Ohta, I. Uchiyama et al., "Whole genome sequencing of meticillin-resistant *Staphylococcus aureus*," *Lancet (London, England)*, vol. 357, no. 9264, pp. 1225–1240, 2001.
- [10] M. Kellis, B. W. Birren, and E. S. Lander, "Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*," *Nature*, vol. 428, no. 6983, pp. 617–624, 2004.
- [11] K. L. Adams and J. F. Wendel, "Polyploidy and genome evolution in plants," *Current Opinion in Plant Biology*, vol. 8, no. 2, pp. 135–141, 2005.
- [12] B. K. Mable, M. A. Alexandrou, and M. I. Taylor, "Genome duplication in amphibians and fish : an extended synthesis," *Journal of Zoology*, vol. 284, no. 3, pp. 151–182, 2011.
- [13] S. M. K. Glasauer and S. C. F. Neuhauss, "Whole-genome duplication in teleost fishes and its evolutionary consequences," *Molecular Genetics and Genomics: MGG*, vol. 289, no. 6, pp. 1045–1060, 2014.
- [14] V. Ravi and B. Venkatesh, "Rapidly evolving fish genomes and teleost diversity," *Current Opinion in Genetics & Development*, vol. 18, no. 6, pp. 544–550, 2008.

- [15] A. Christoffels, E. G. L. Koh, J. Chia, S. Brenner, S. Aparicio, and B. Venkatesh, "Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes," *Molecular Biology and Evolution*, vol. 21, no. 6, pp. 1146–1151, 2004.
- [16] M. Robinson-Rechavi, O. Marchand, H. Escriva, and V. Laudet, "An ancestral whole-genome duplication may not have been responsible for the abundance of duplicated fish genes," *Current Biology: CB*, vol. 11, no. 12, pp. R458–R459, 2001.
- [17] J. S. Taylor, I. Braasch, T. Frickey, A. Meyer, and Y. Van de Peer, "Genome duplication, a trait shared by 22000 species of ray-finned fish," *Genome Research*, vol. 13, no. 3, pp. 382–390, 2003.
- [18] J. S. Taylor and J. Raes, "Duplication and divergence : the evolution of new genes and old ideas," *Annual Review of Genetics*, vol. 38, no. 1, pp. 615–643, 2004.
- [19] Y. Van de Peer, S. Maere, and A. Meyer, "The evolutionary significance of ancient genome duplications," *Genetics*, vol. 10, no. 10, pp. 725–732, 2009.
- [20] Y. Van de Peer, E. Mizrahi, and K. Marchal, "The evolutionary significance of polyploidy," *Genetics*, vol. 18, no. 7, pp. 411–424, 2017.
- [21] S. Maere, S. De Bodt, J. Raes et al., "Modeling gene and genome duplications in eukaryotes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 15, pp. 5454–5459, 2005.
- [22] D. Sankoff, C. Zheng, and Q. Zhu, "The collapse of gene complement following whole genome duplication," *BMC Genomics*, vol. 11, no. 1, p. 313, 2010.
- [23] C. Berthelot, F. Brunet, D. Chalopin et al., "The rainbow trout genome provides novel insights into evolution after whole-genome duplication in vertebrates," *Nature Communications*, vol. 5, no. 1, p. 3657, 2014.
- [24] G. Conant and K. Wolfe, "Turning a hobby into a job : how duplicated genes find new functions," *Nature reviews Genetics*, vol. 9, pp. 938–950, 2008.
- [25] A. Grandchamp, B. Piégu, and P. Monget, "Genes encoding teleost fish ligands and associated receptors remained in duplicate more frequently than the rest of the genome," *Genome Biology and Evolution*, vol. 11, no. 5, pp. 1451–1462, 2019.
- [26] A. Nag, S. Vigneau, V. Savova, L. M. Zwemer, and A. A. Gimelbrant, "Chromatin signature identifies monoallelic gene expression across mammalian cell types," *G3 (Bethesda, Md.)*, vol. 5, no. 8, pp. 1713–1720, 2015.
- [27] L. Carrel and H. F. Willard, "X-inactivation profile reveals extensive variability in X-linked gene expression in females," *Nature*, vol. 434, no. 7031, pp. 400–404, 2005.

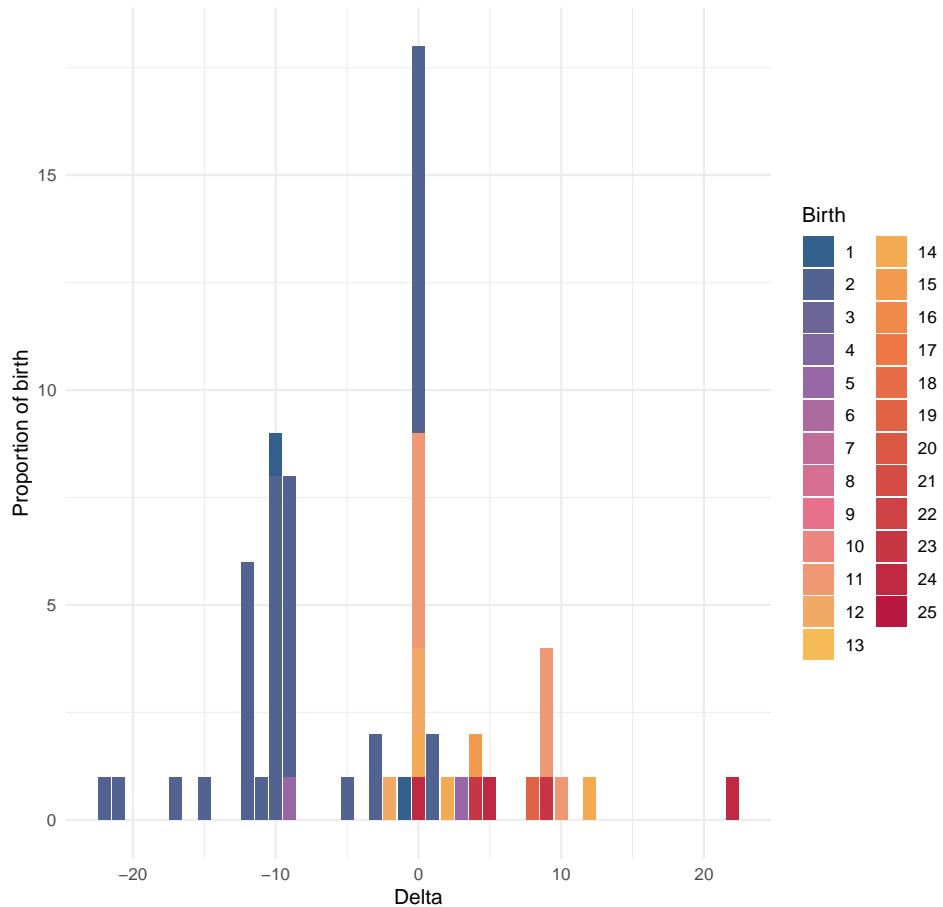
## 6.2 Article 1 - Suppl. Data 1

**Suppl. Data 1 - Distribution of deltas of node of birth of genes encoding proteins involved in all the pathways, according to the node of birth of each gene**

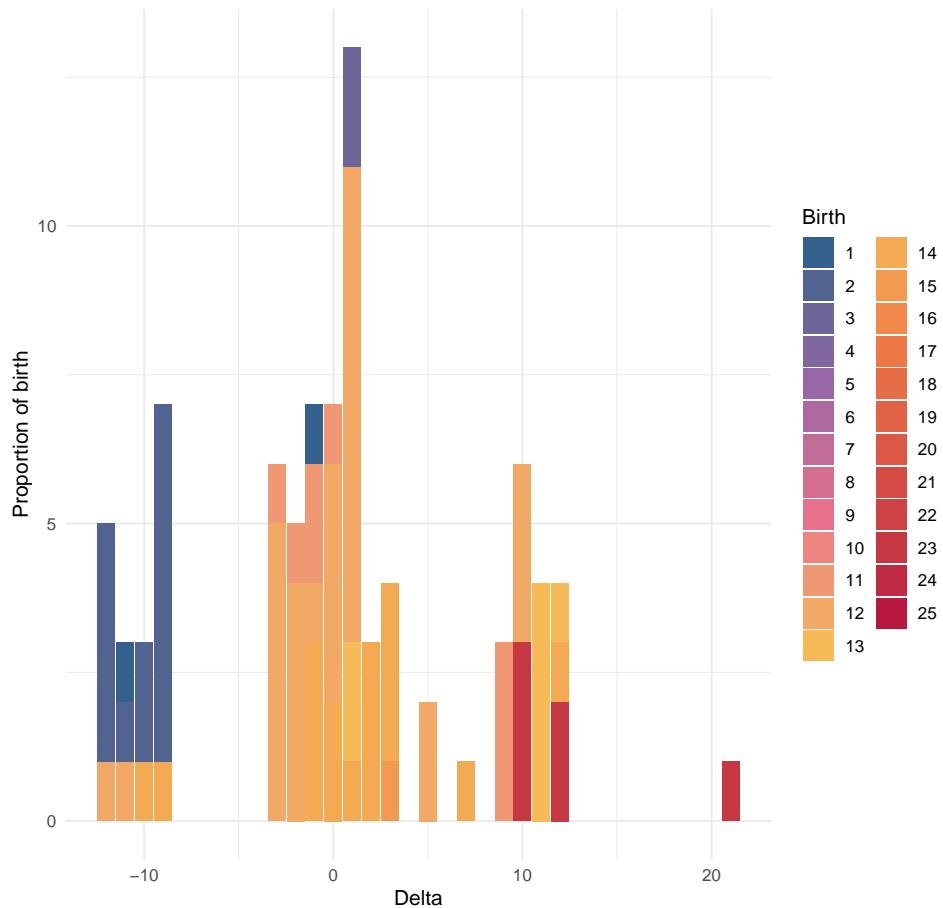
Legend : Deltas are calculated via clade of birth rank for the A gene - clade of birth rank for the B gene for the interaction A → B. For example, if gene A was born at the blue clade (clade 1), and the clade of birth of B is 11, the A → B delta is -10. Moreover, in this case, it is a backward relationship, because A was born before B.

The pathways are in the following order : p53, AGE-RAGE, Adipocytokine, Estrogen, Glucagon, GnRH, Insulin, Ovarian steroidogenesis, Oxytocin, PPAR, Prolactin, Relaxin, Thyroid hormone, B cell receptor, C-type lectin receptor, Chemokine, FC epsilon RI, IL-17, NOD-like receptor, RIG-I-like receptor, T cell receptor, Toll-like receptor, Neurotrophin, AMPK, Apelin, Calcium, cAMP, cGMP-PKG, ErbB, FoxO, Hedgehog, HIF-1, Hippo, JAK-STAT, MAPK, mTOR, NF-Kappa B, Notch, Phospholipase D, PI3K-Akt, Rap1, Ras, Sphingolipid, TGF-Beta, TNF , VEGF, Wnt.

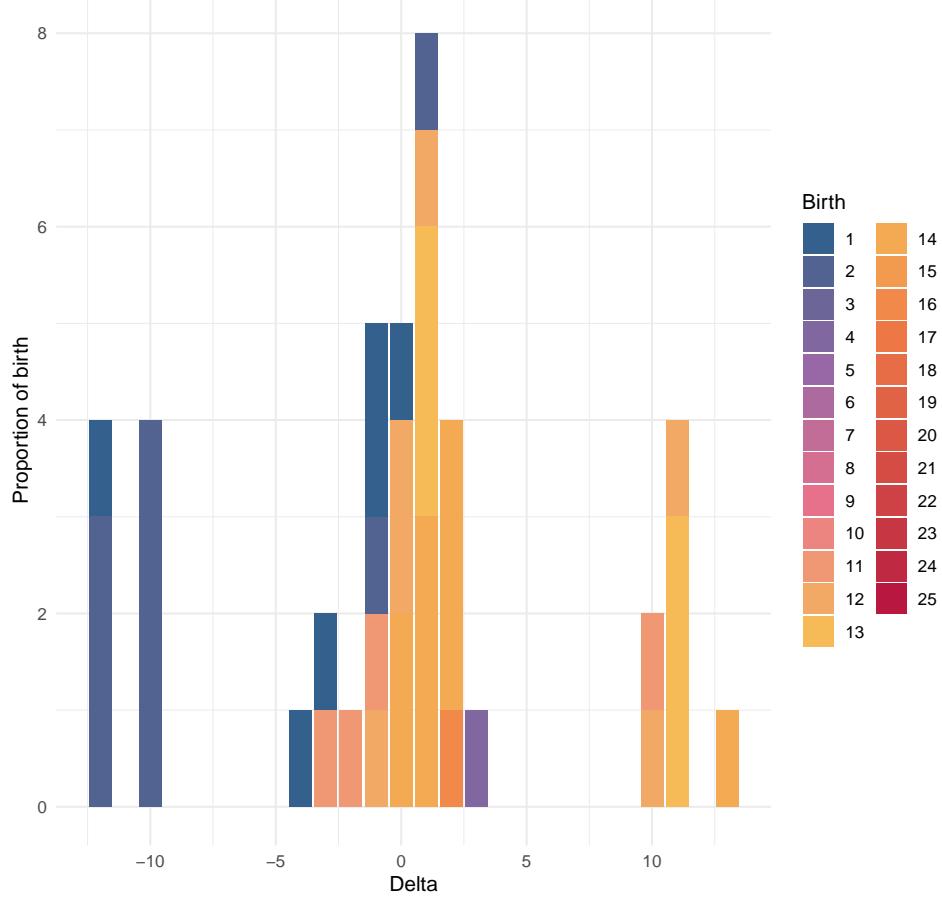
Distribution of birth moments by delta for p53



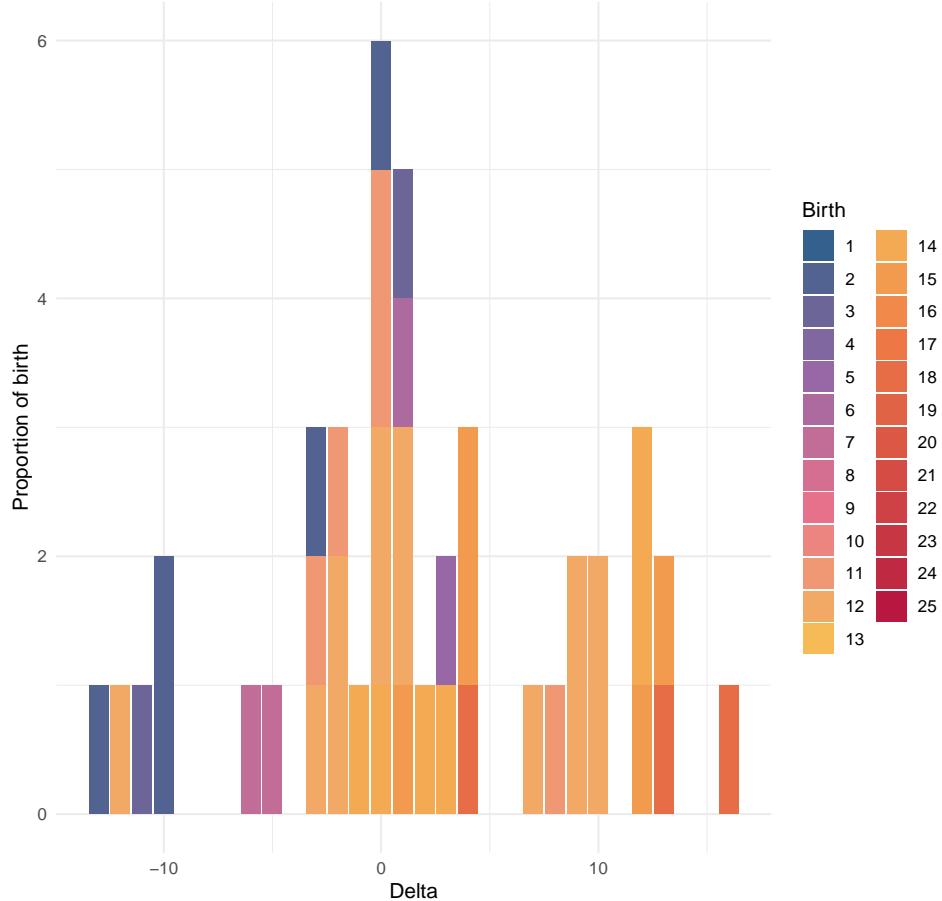
Distribution of birth moments by delta for AGE-RAGE



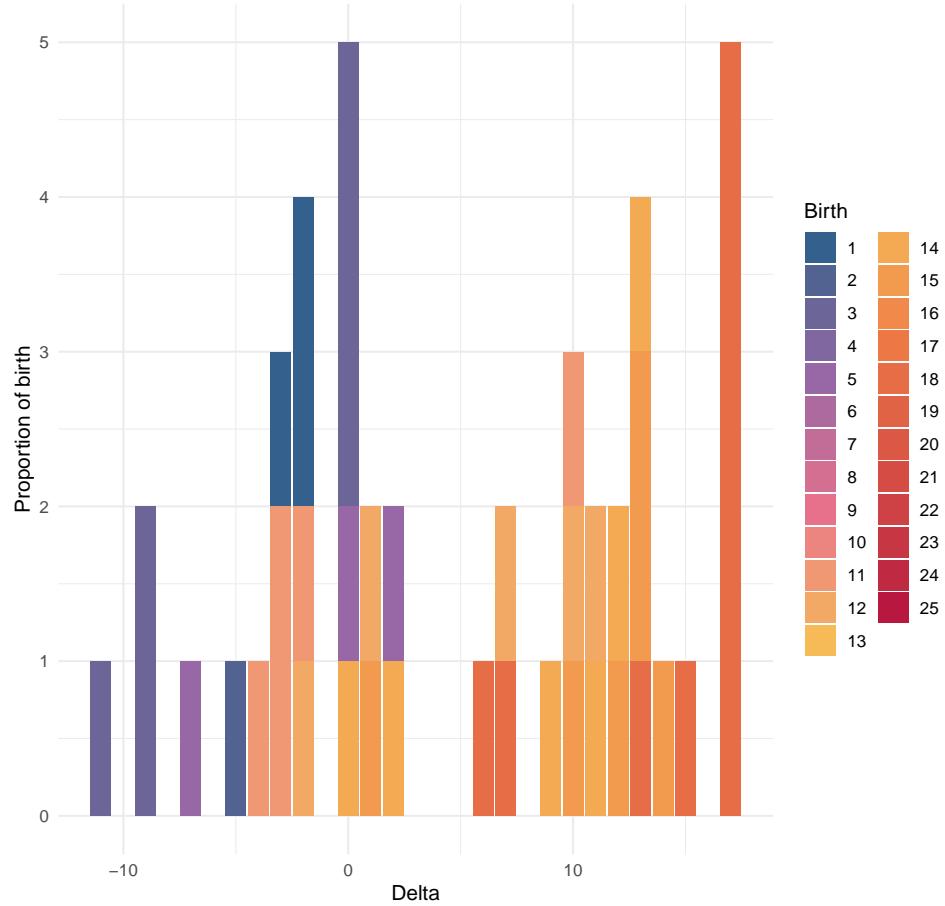
Distribution of birth moments by delta for Adipocytokine



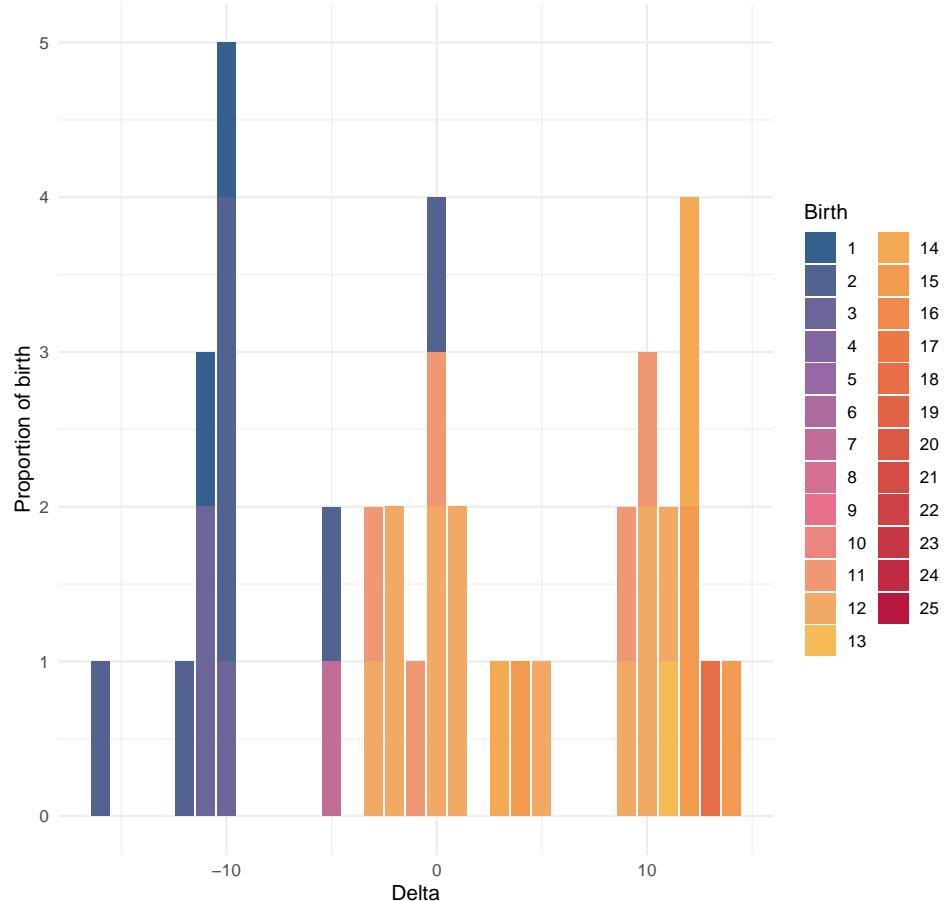
Distribution of birth moments by delta for Estrogen



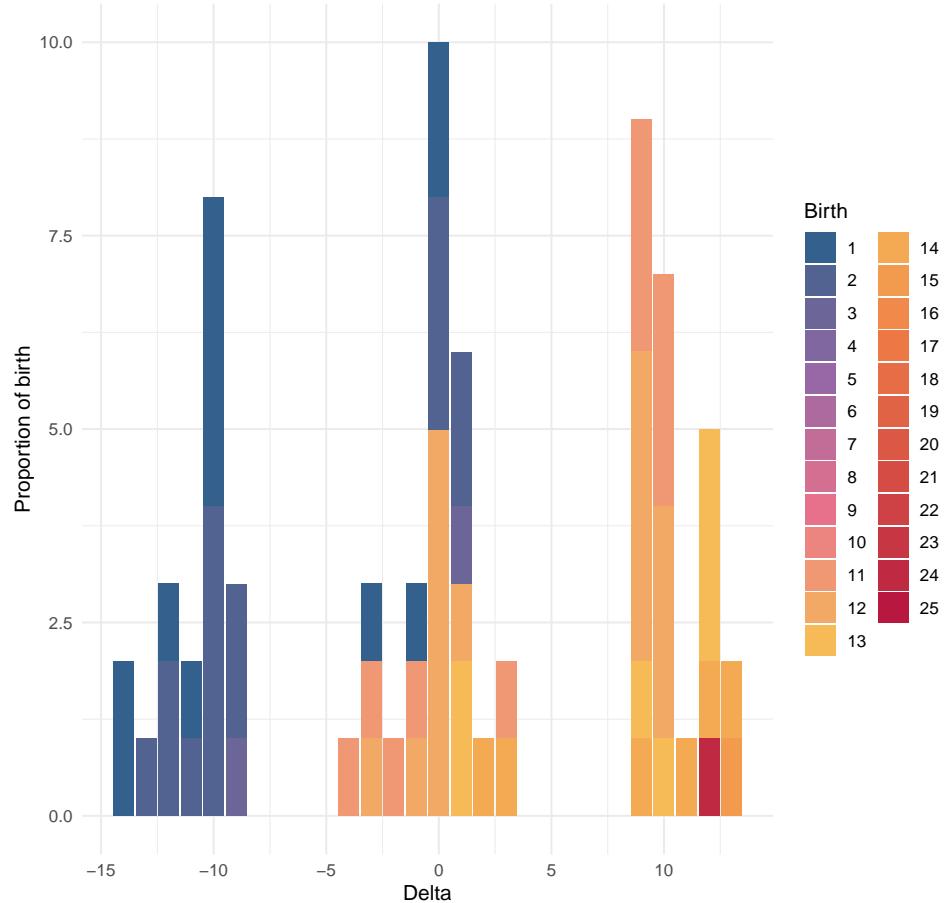
Distribution of birth moments by delta for Glucagon



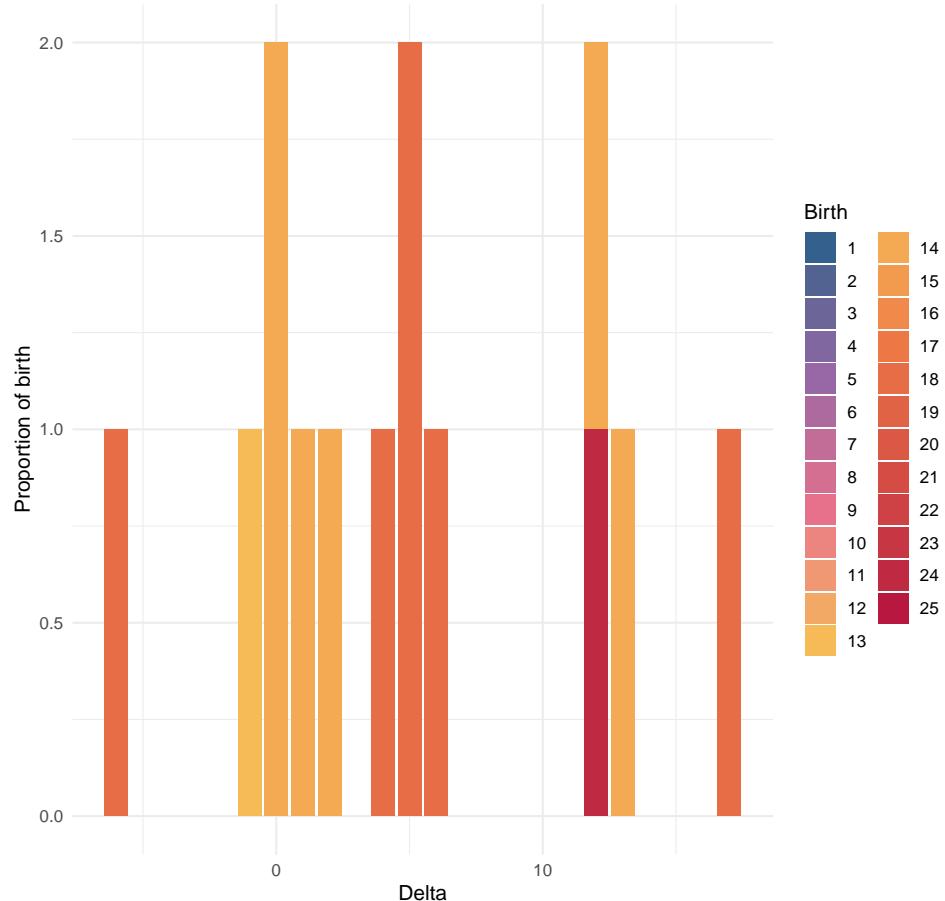
Distribution of birth moments by delta for GnRH



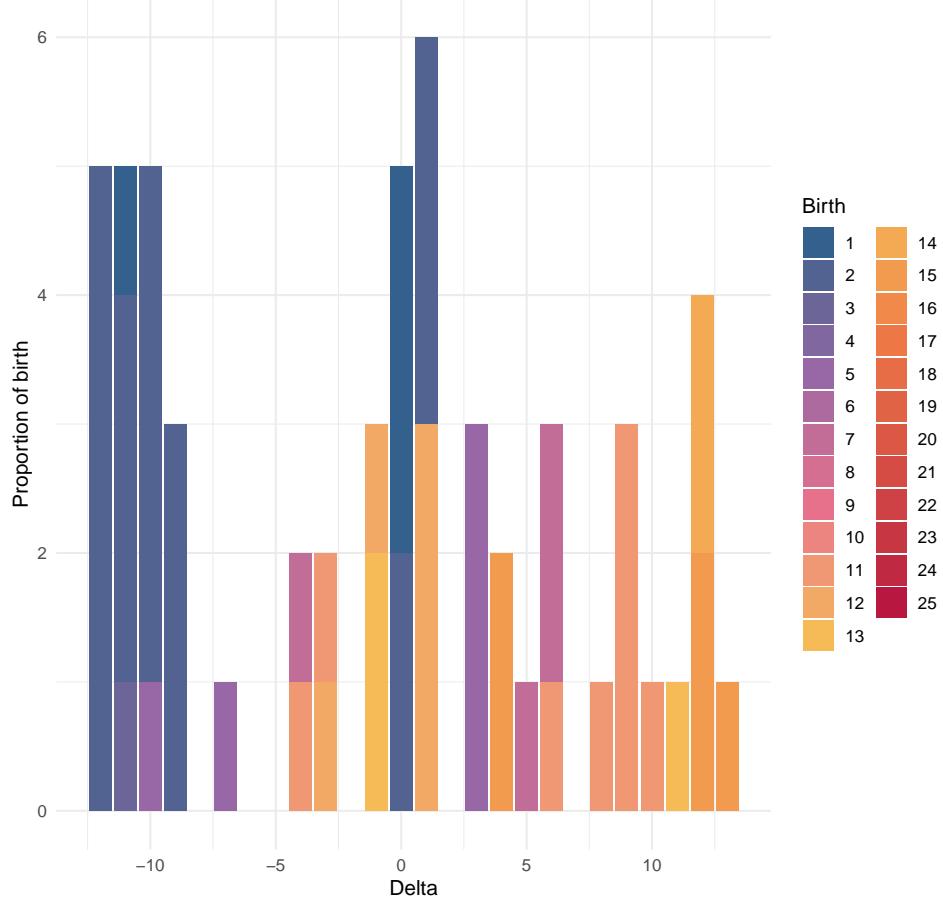
Distribution of birth moments by delta for Insulin



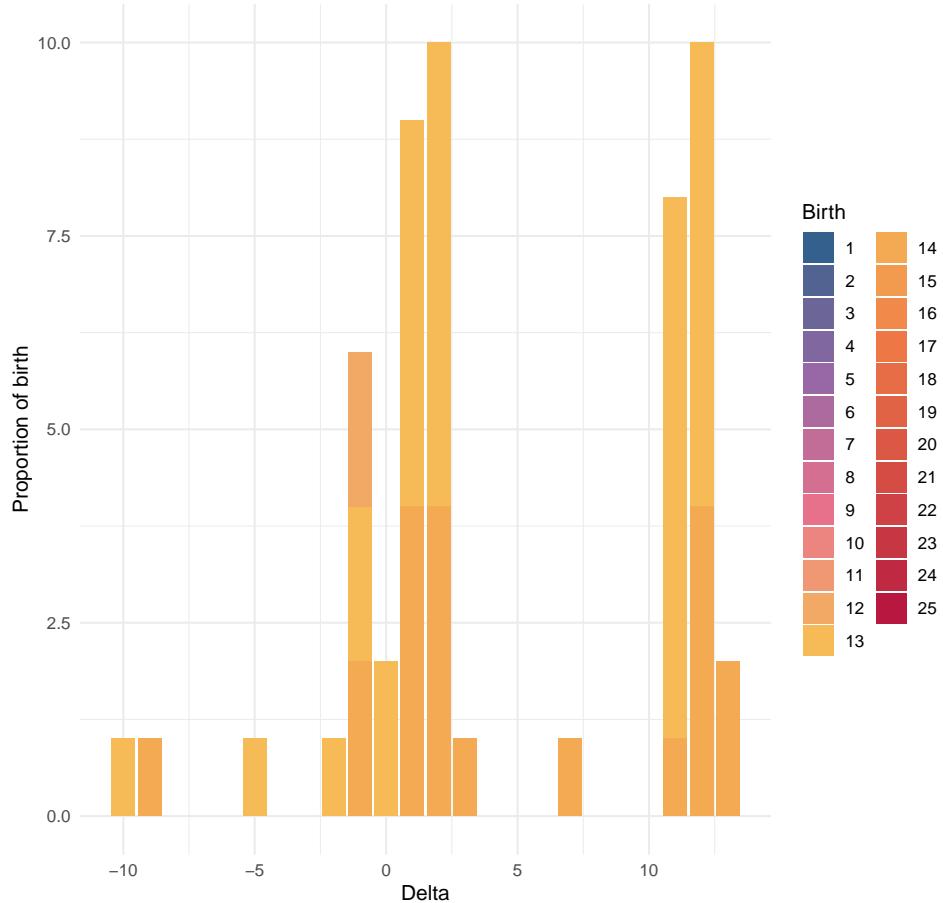
Distribution of birth moments by delta for Ovarian steroidogenesis



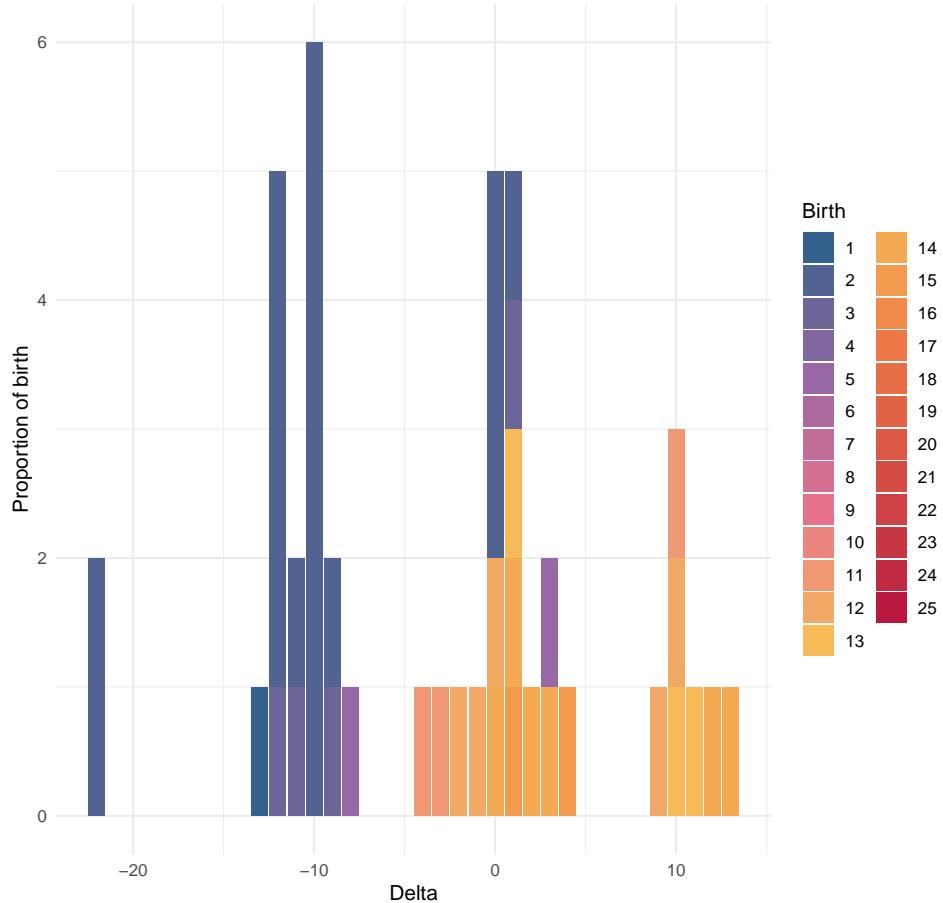
Distribution of birth moments by delta for Oxytocin



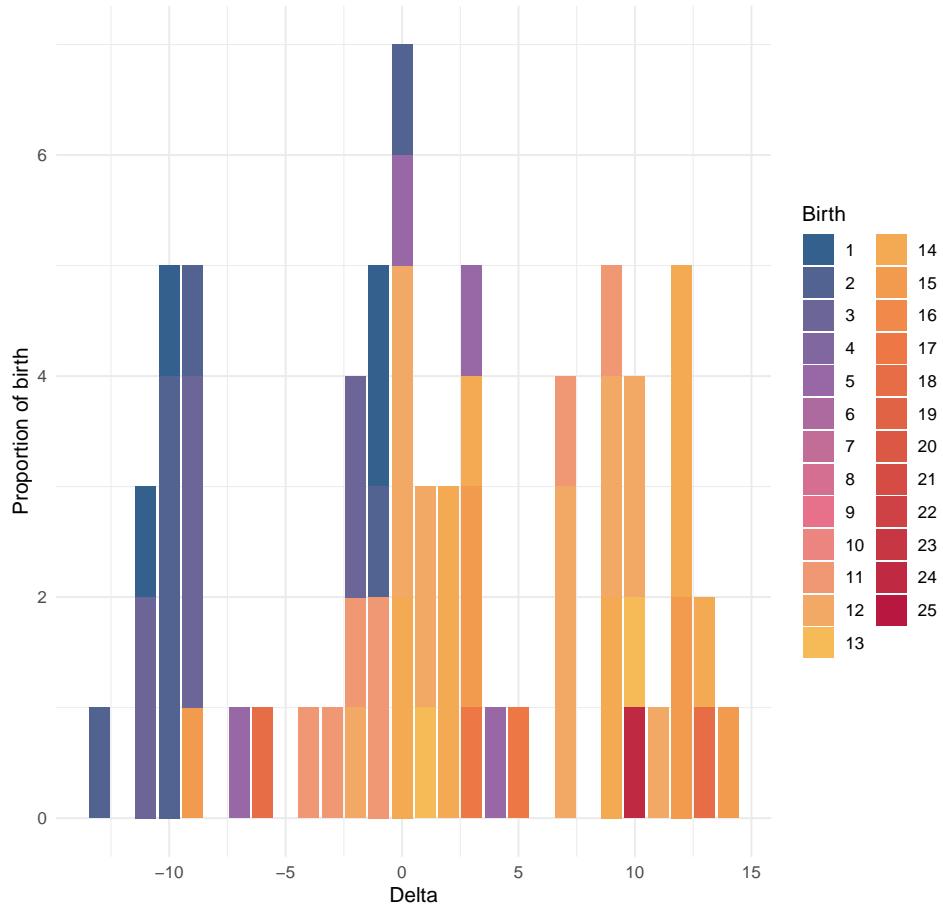
Distribution of birth moments by delta for PPAR



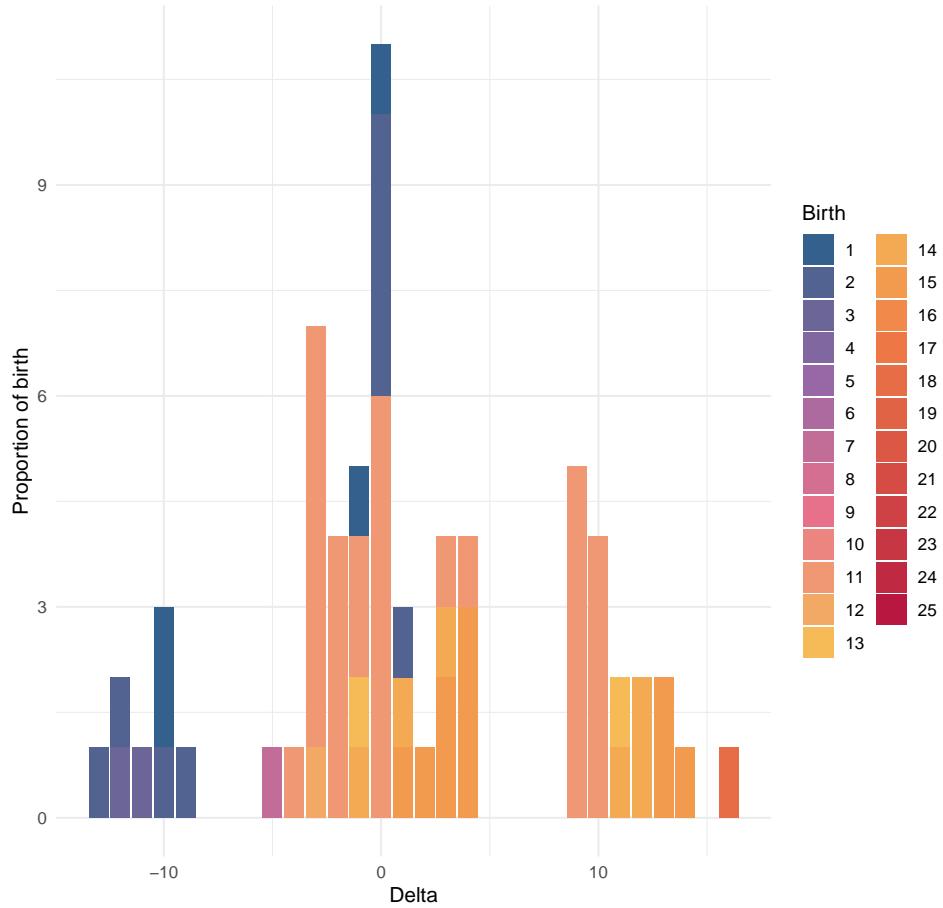
Distribution of birth moments by delta for Prolactin



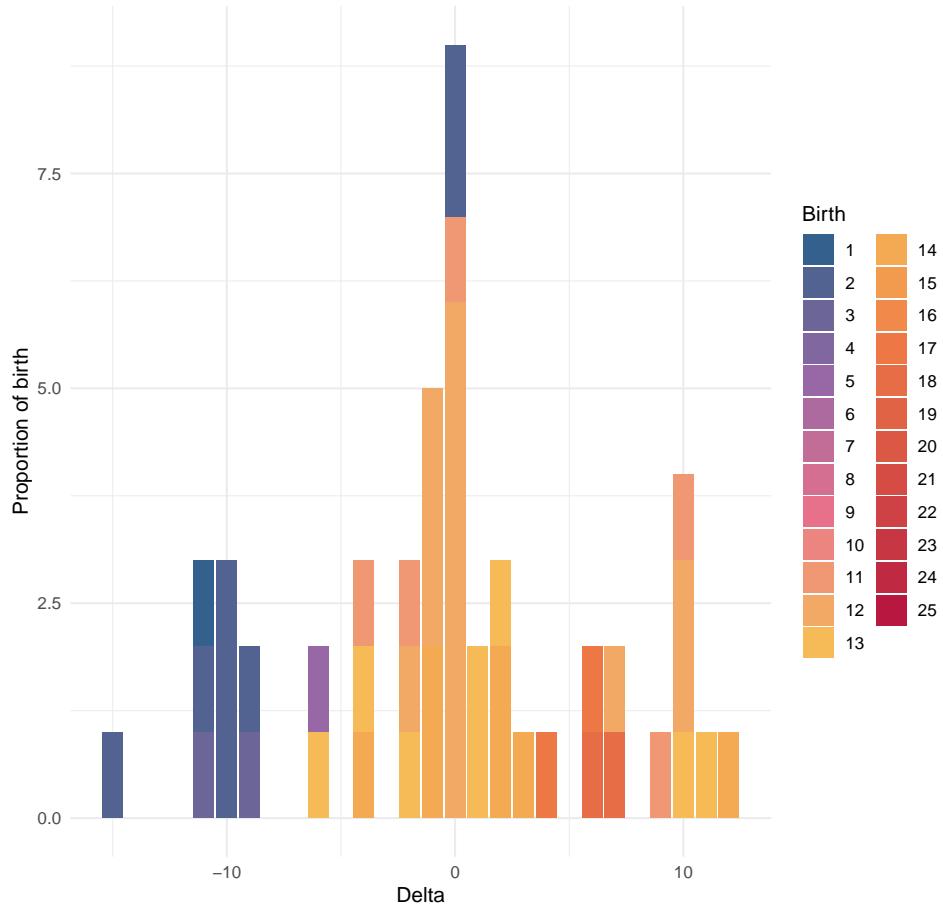
Distribution of birth moments by delta for Relaxin



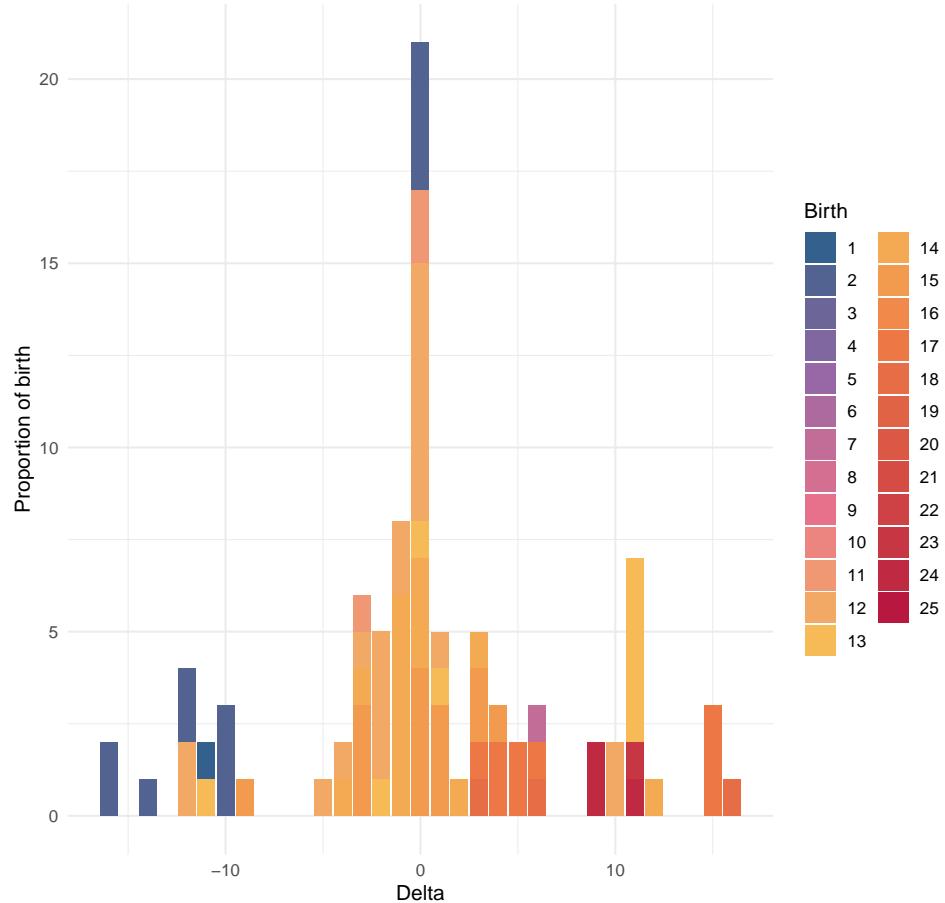
Distribution of birth moments by delta for Thyroid hormone



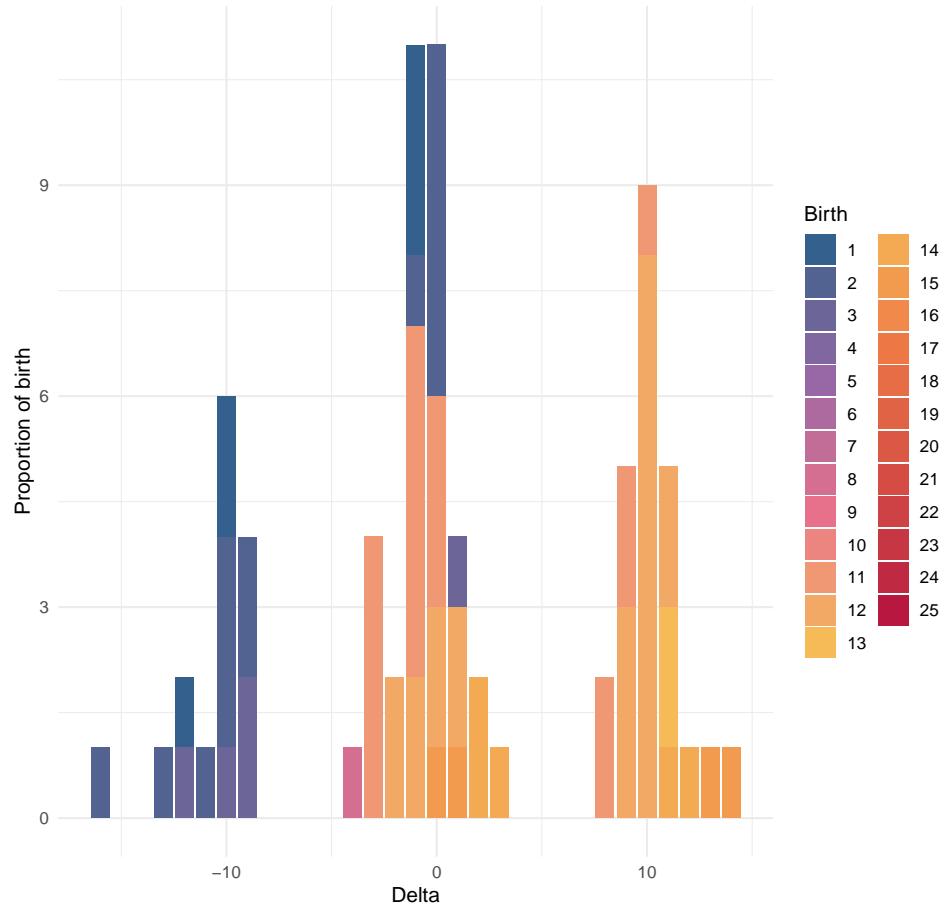
Distribution of birth moments by delta for B cell receptor



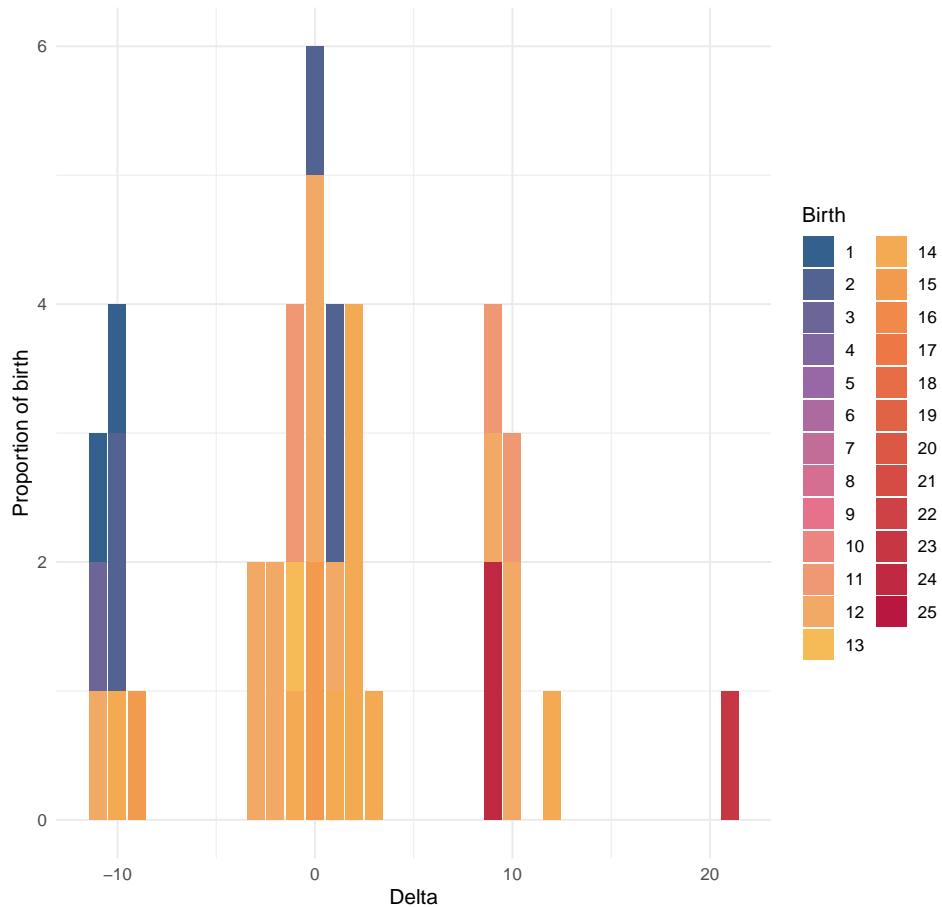
Distribution of birth moments by delta for C-type lectin receptor



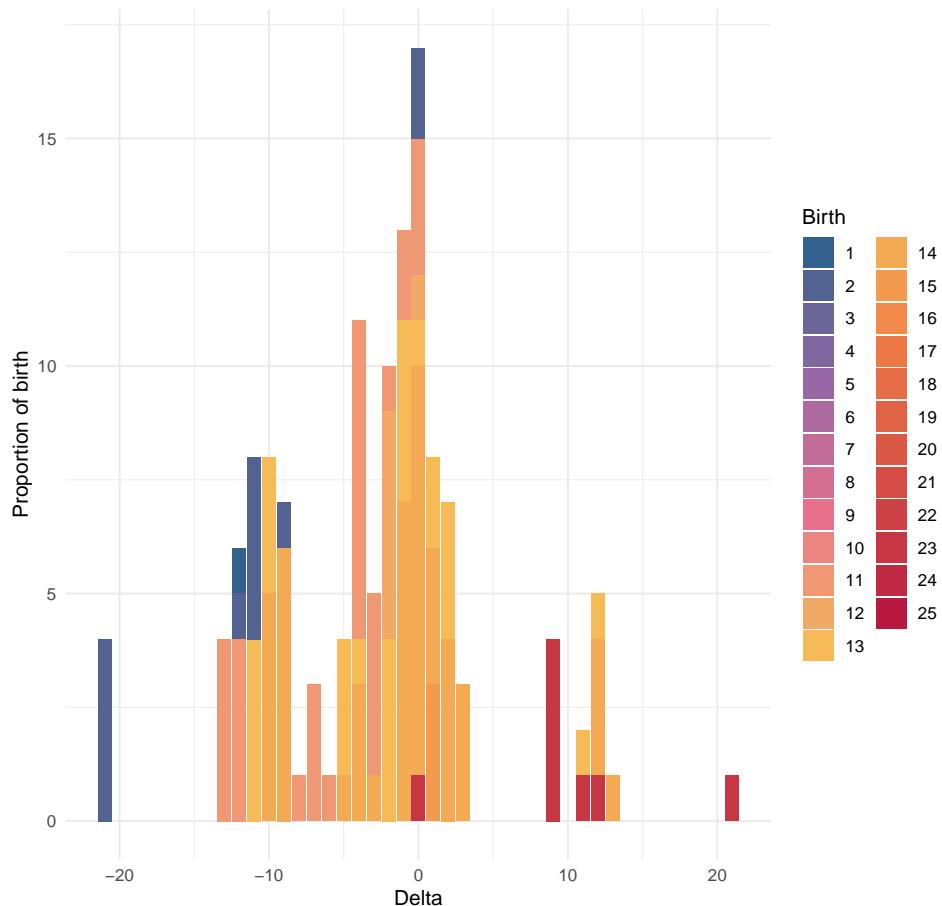
Distribution of birth moments by delta for Chemokine



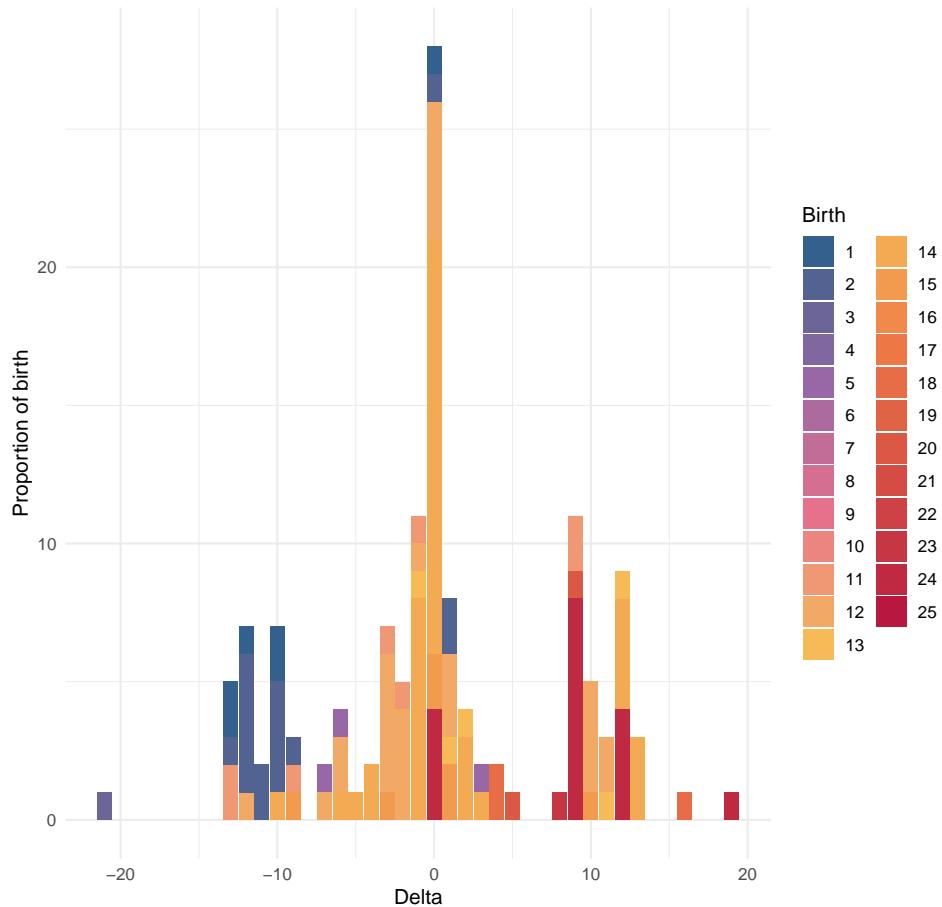
Distribution of birth moments by delta for Fc epsilon RI



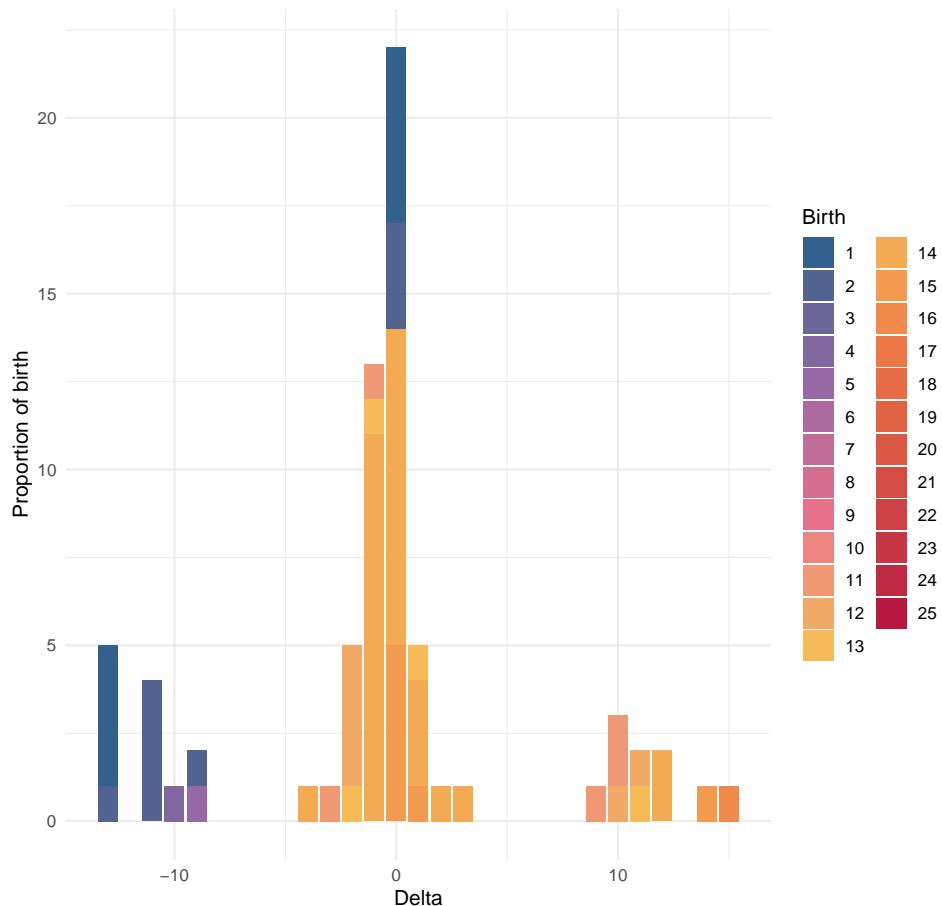
Distribution of birth moments by delta for IL-17



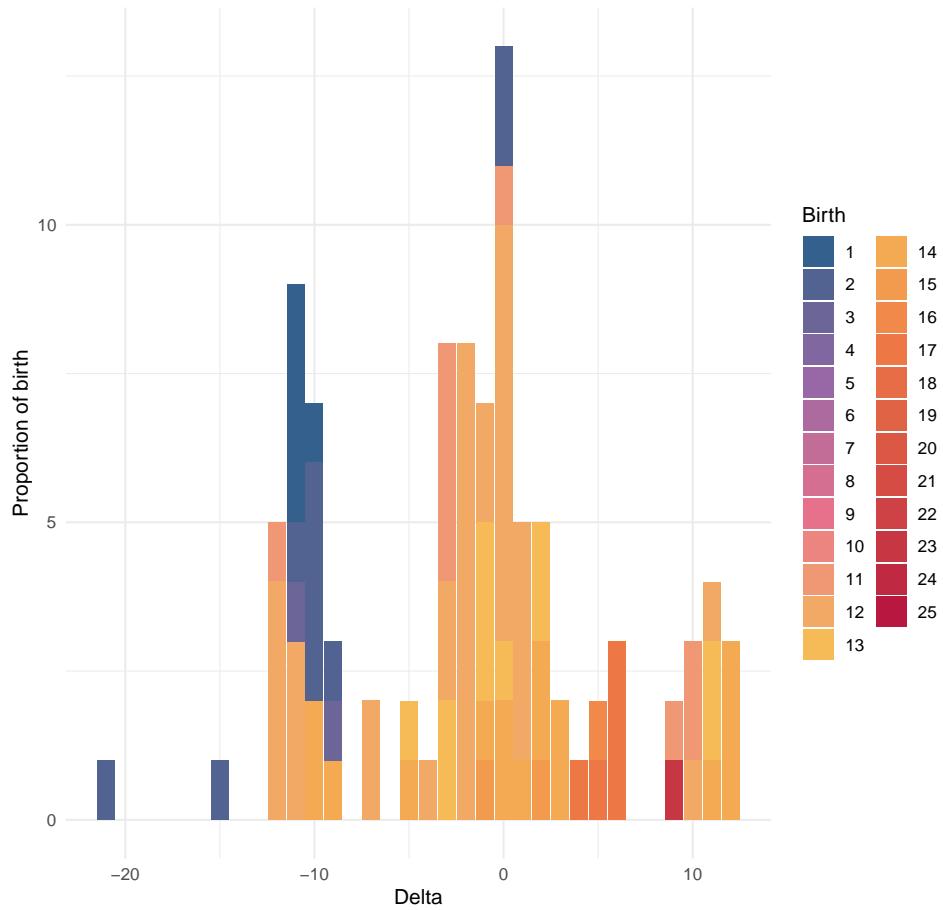
Distribution of birth moments by delta for NOD-like receptor



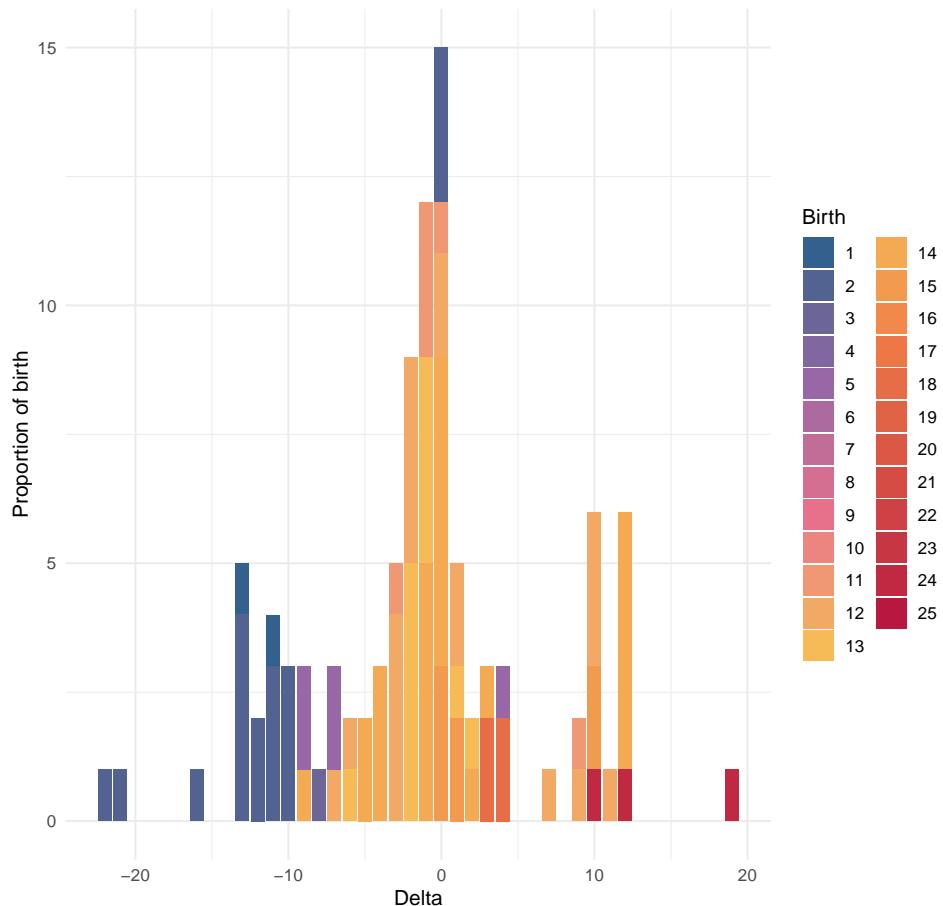
Distribution of birth moments by delta for RIG-I-like receptor



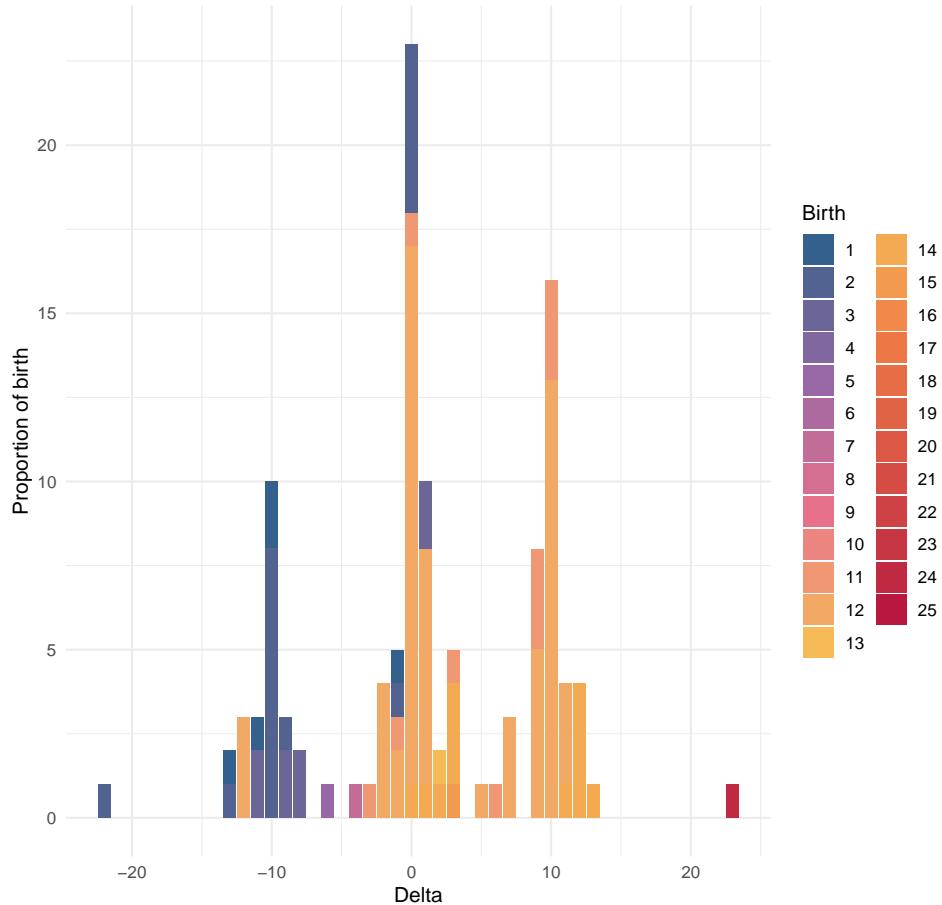
Distribution of birth moments by delta for T cell receptor



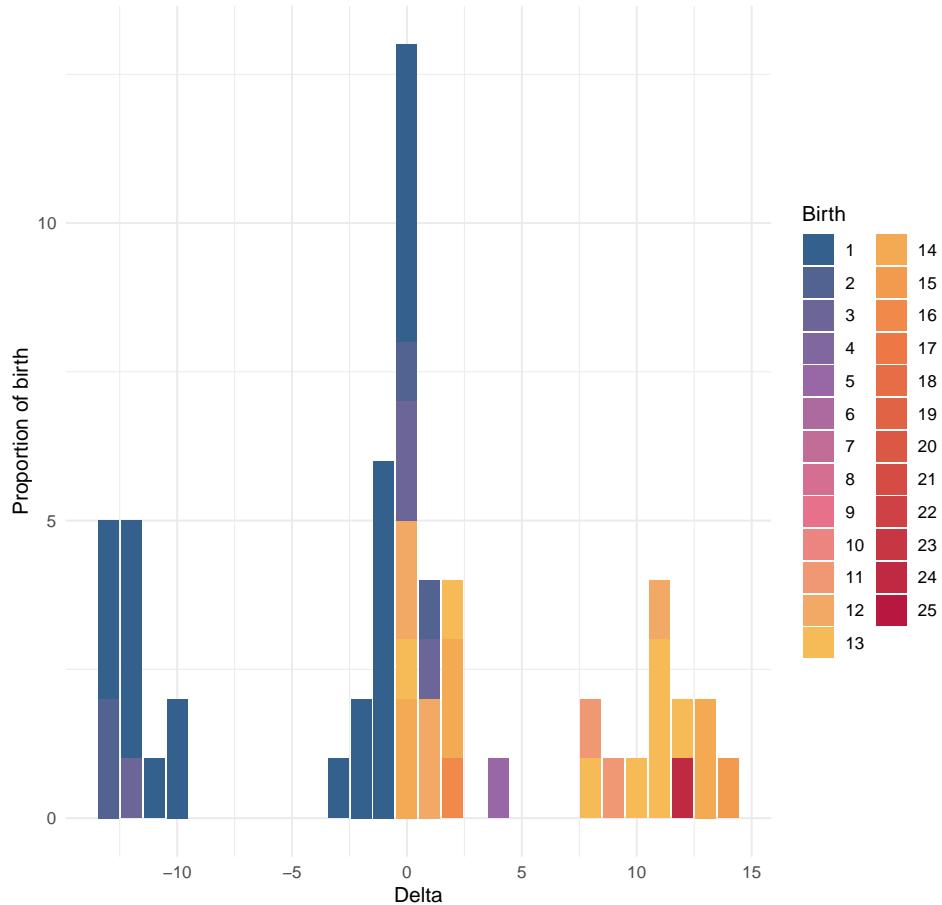
Distribution of birth moments by delta for Toll-like receptor



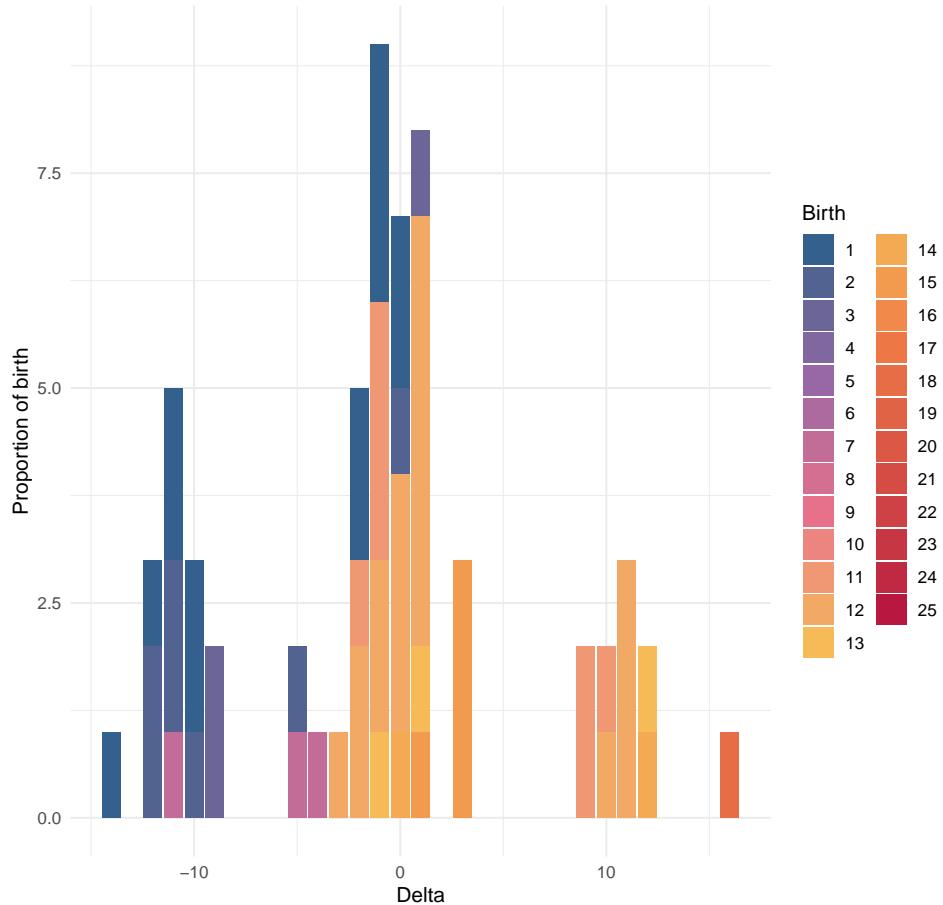
Distribution of birth moments by delta for Neurotrophin



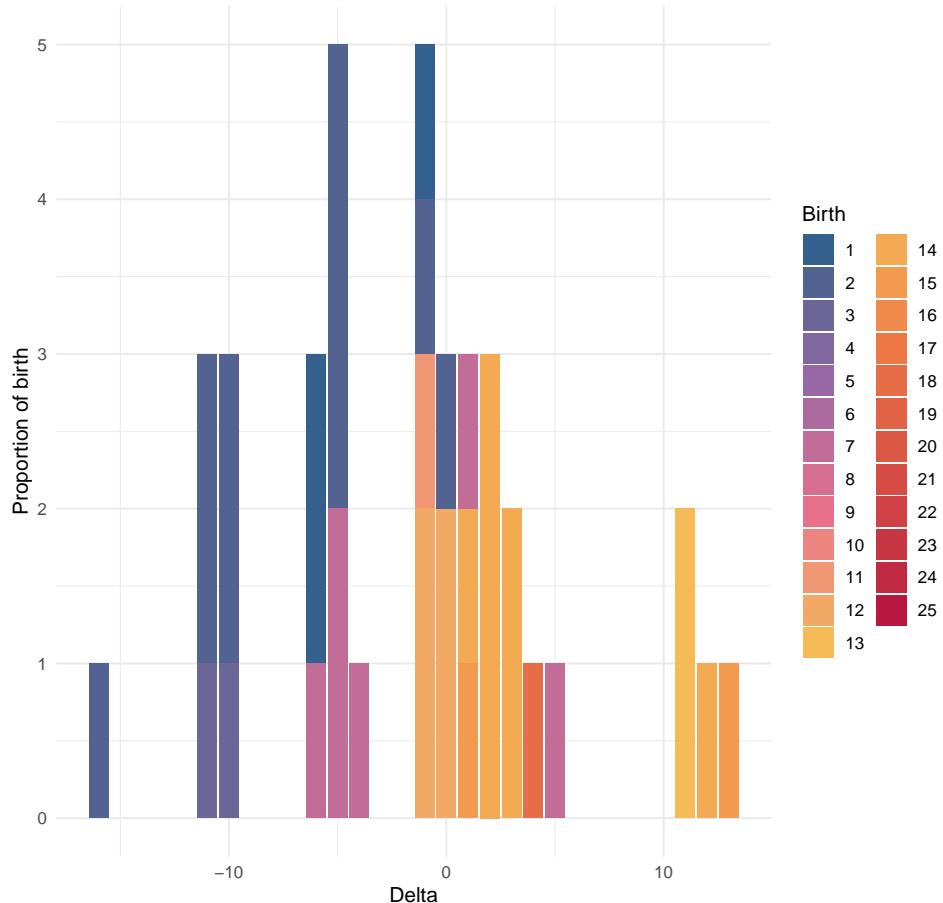
Distribution of birth moments by delta for AMPK



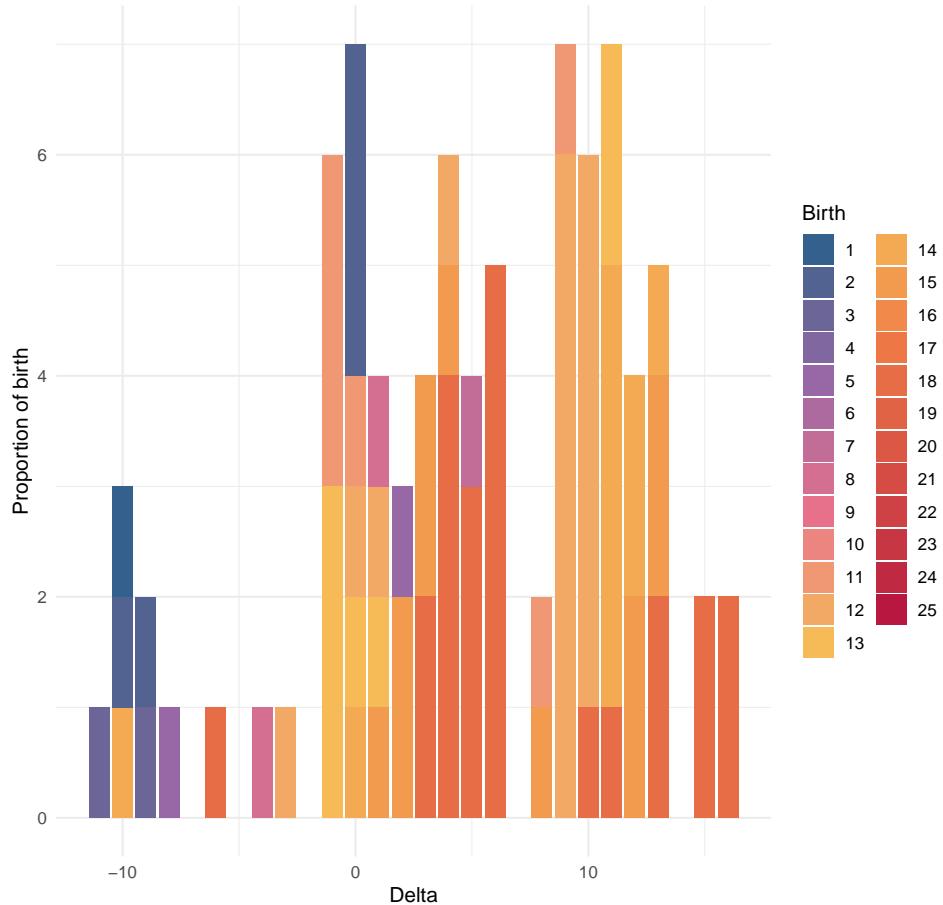
Distribution of birth moments by delta for Apelin



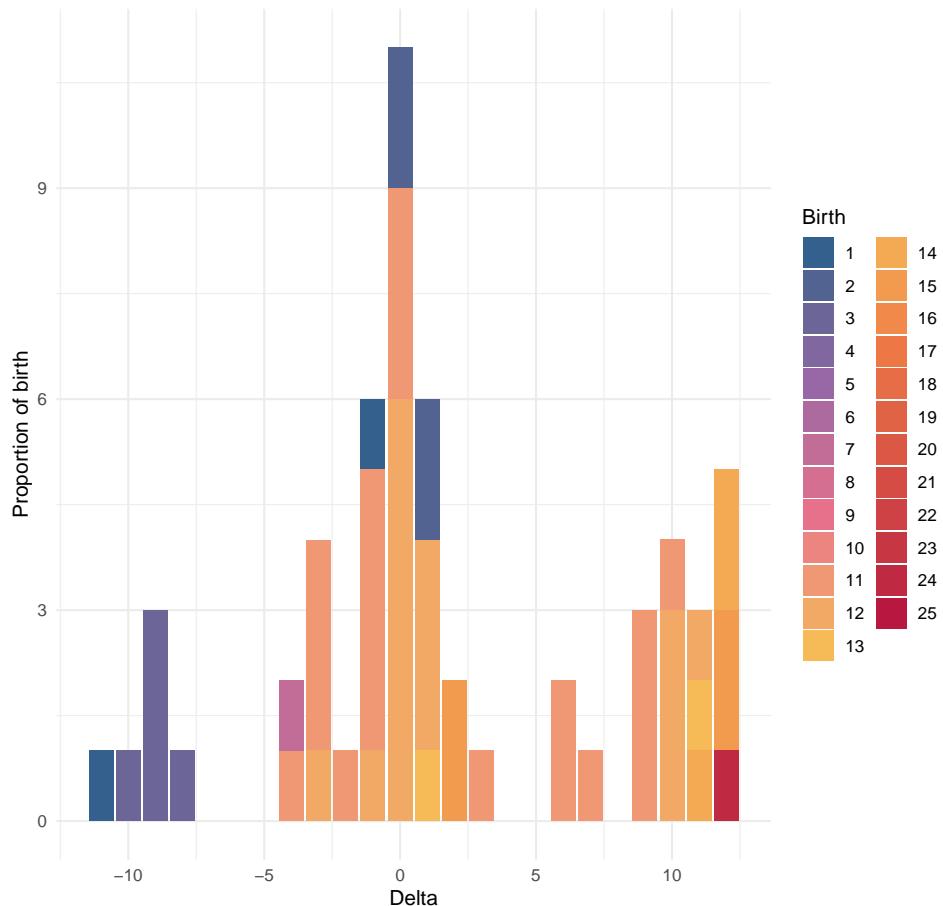
Distribution of birth moments by delta for Calcium

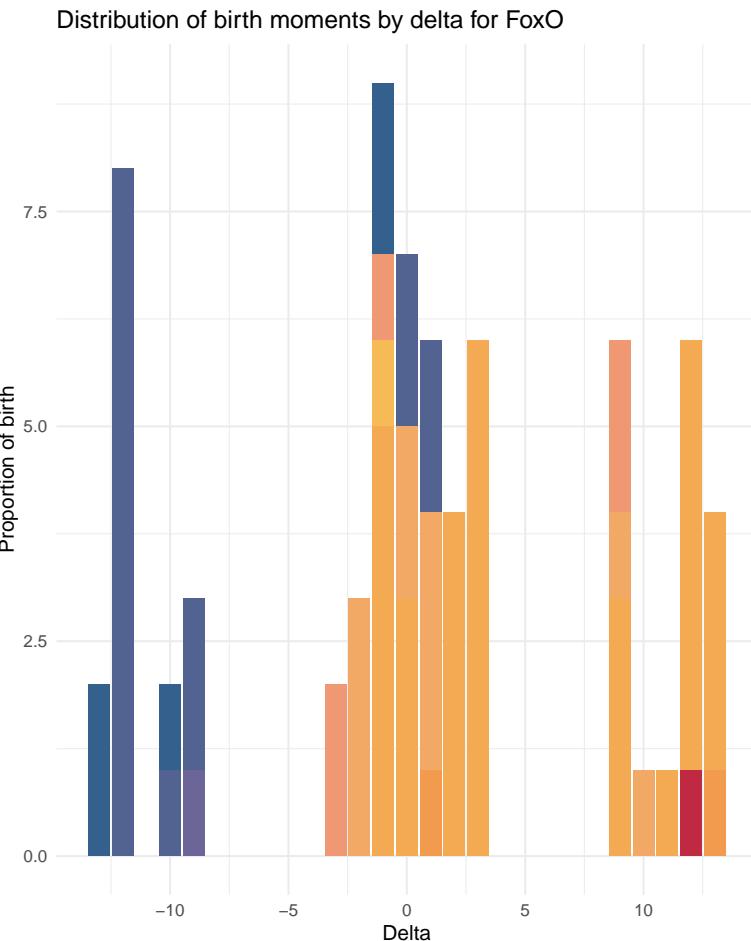
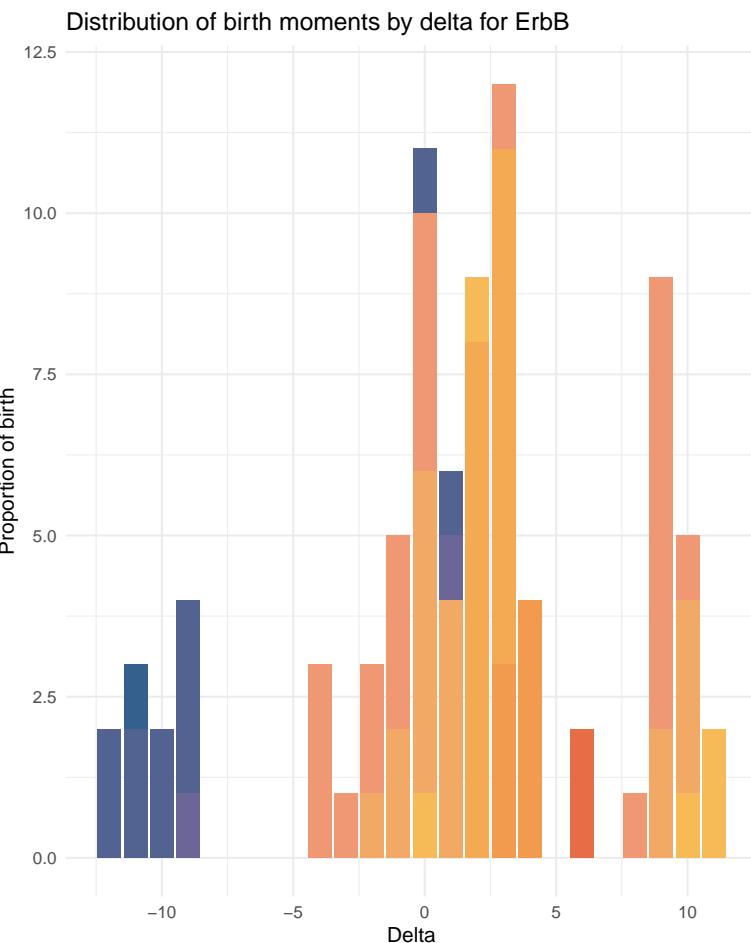


Distribution of birth moments by delta for cAMP

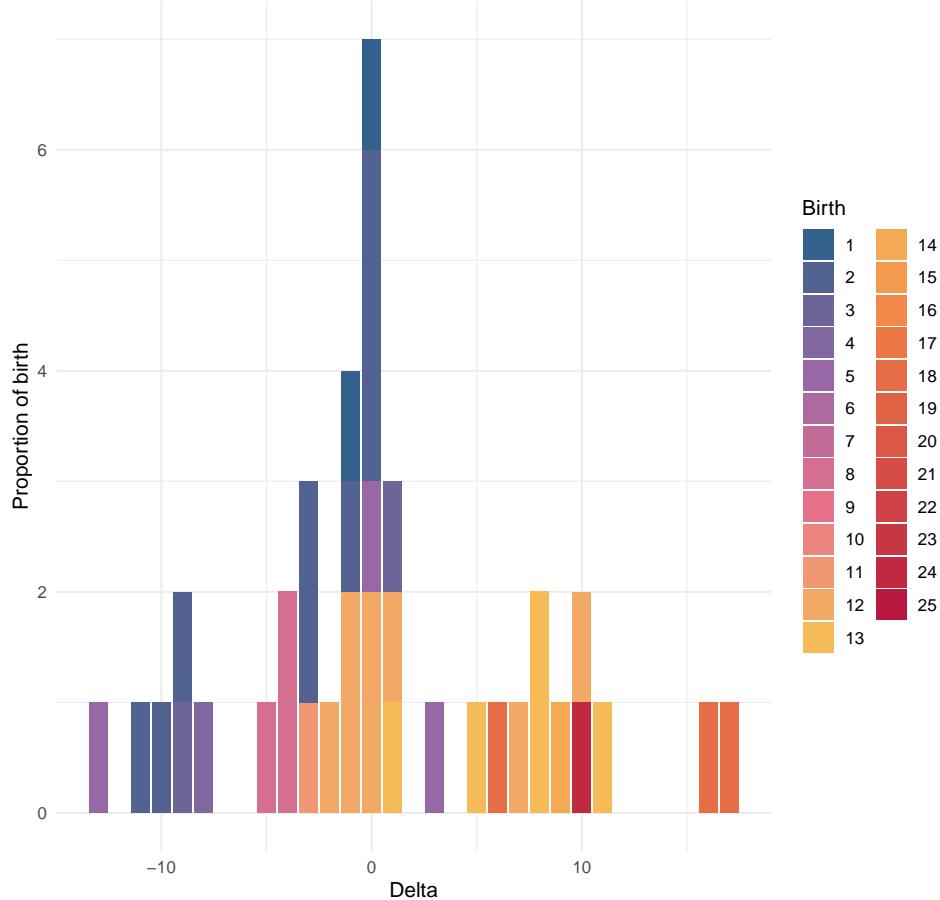


Distribution of birth moments by delta for cGMP–PKG

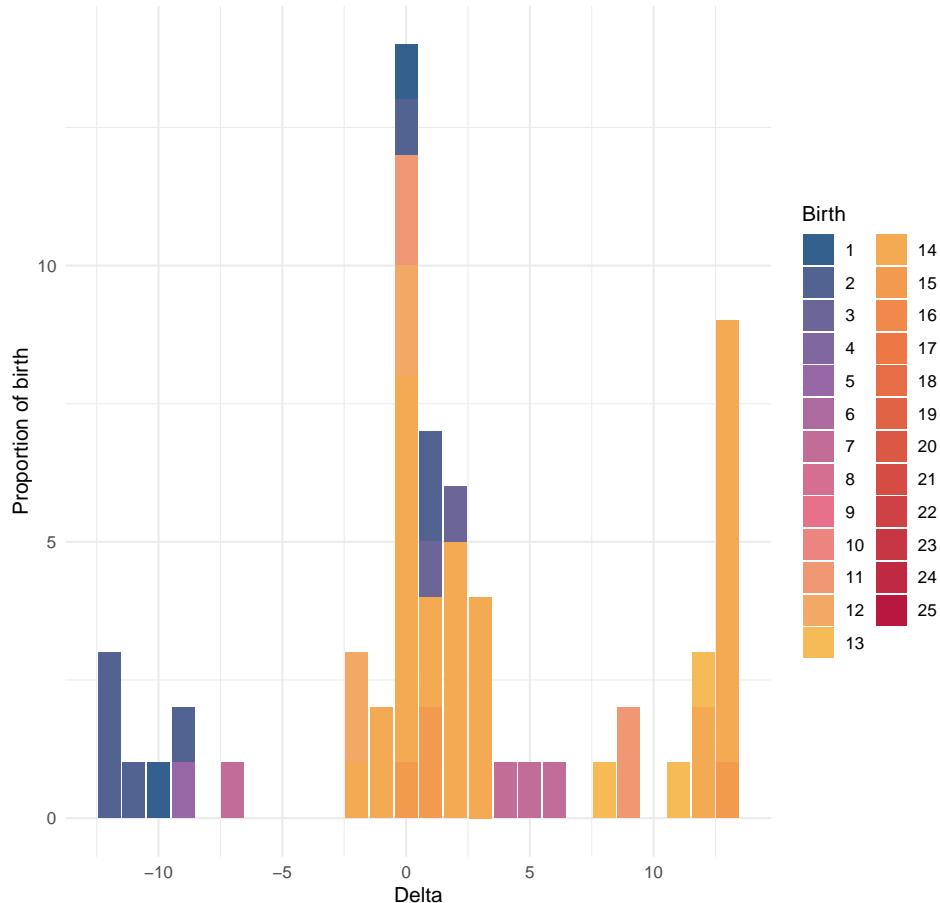




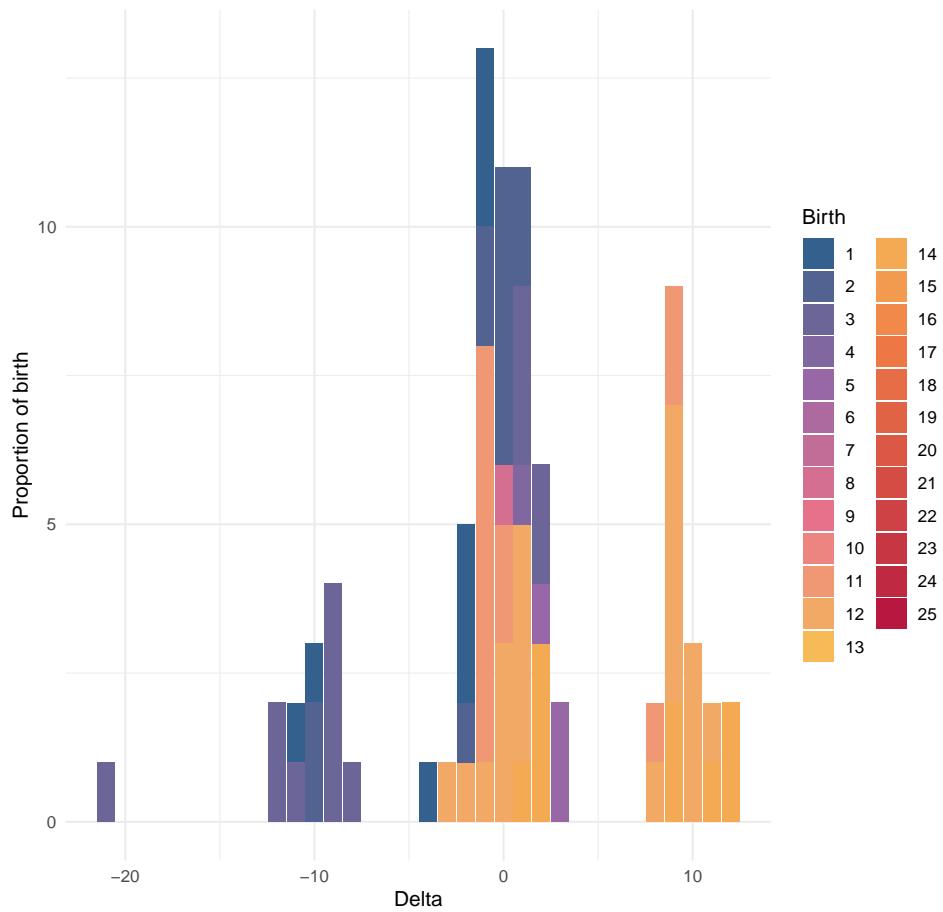
Distribution of birth moments by delta for Hedgehog



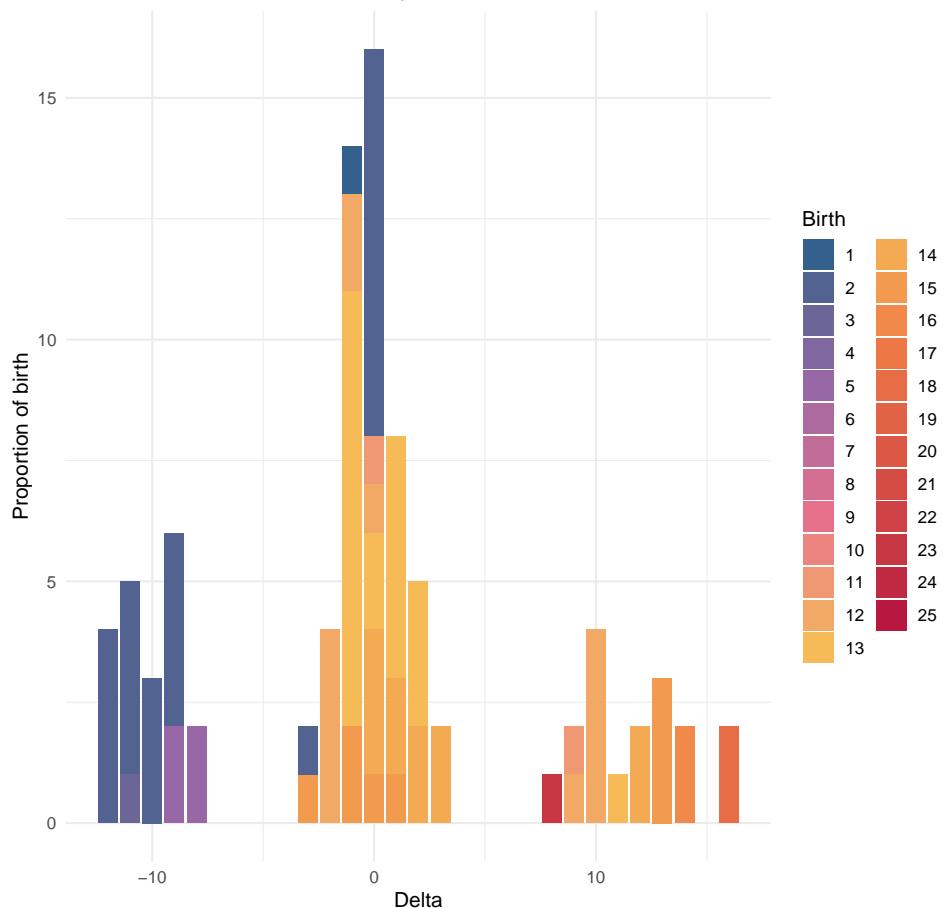
Distribution of birth moments by delta for HIF-1

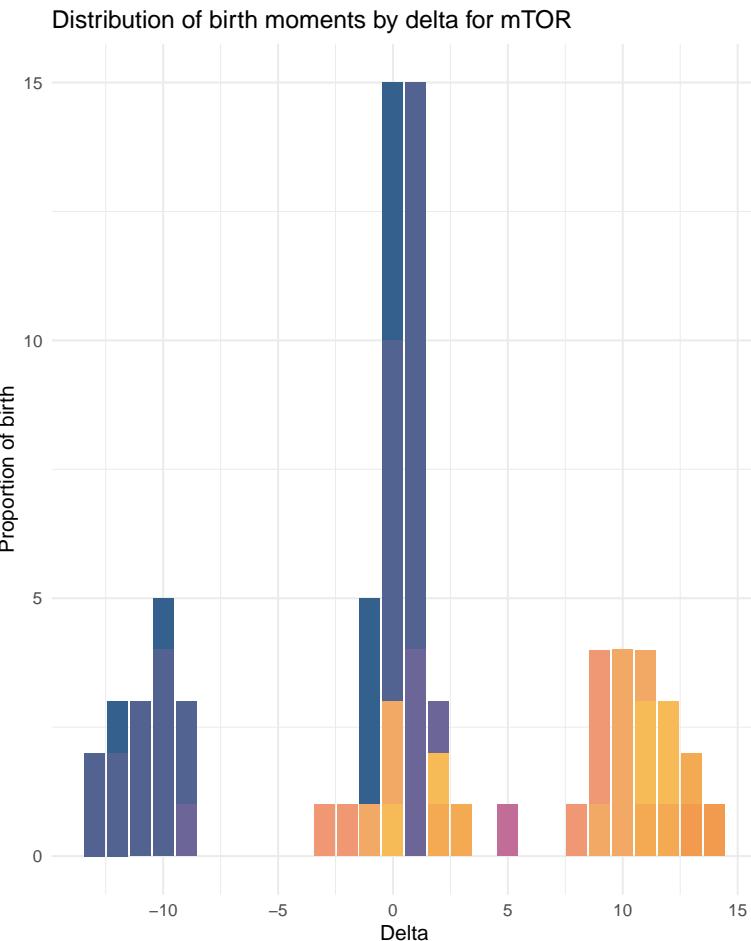
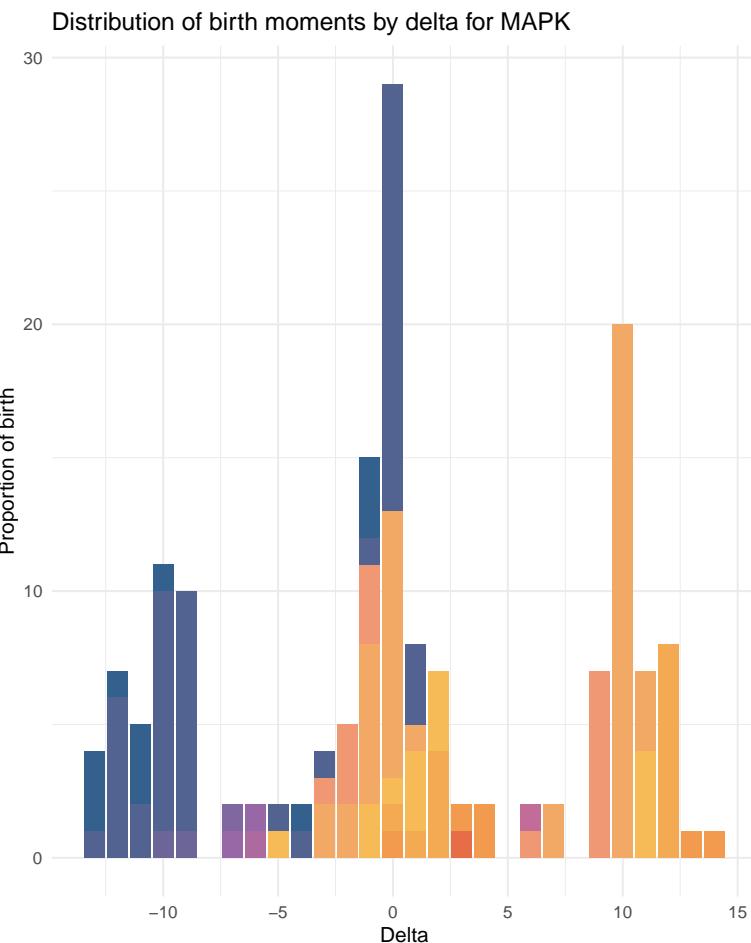


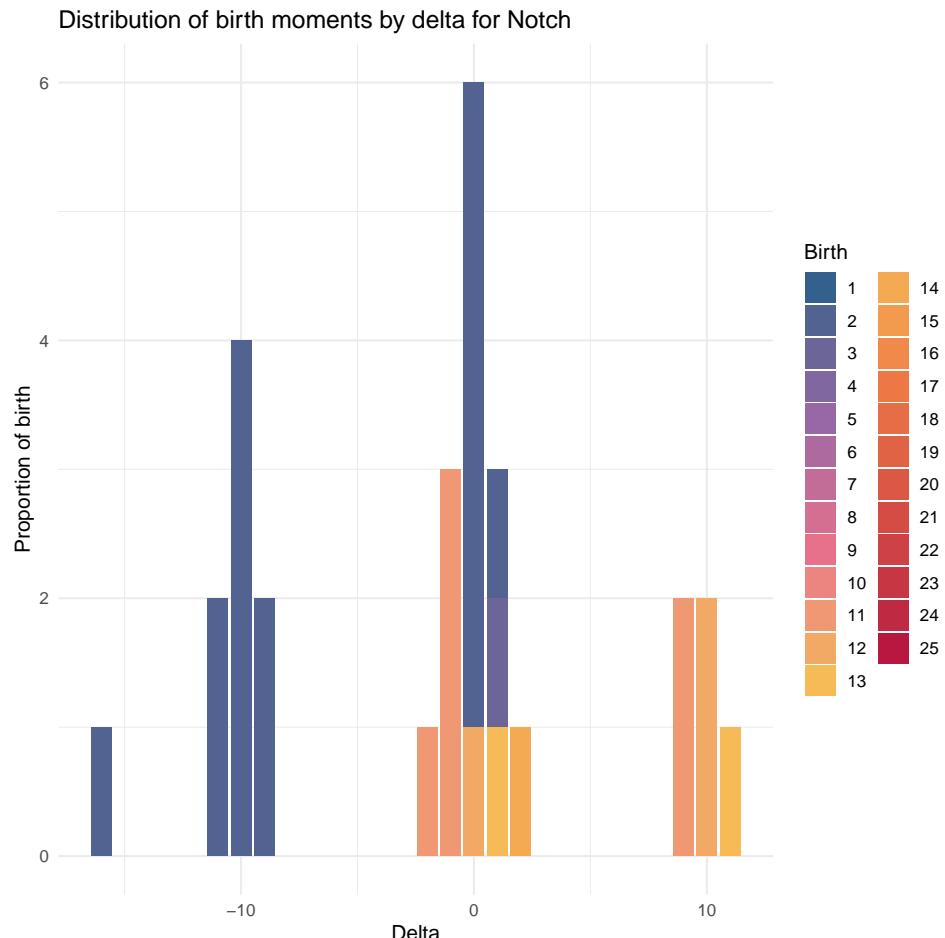
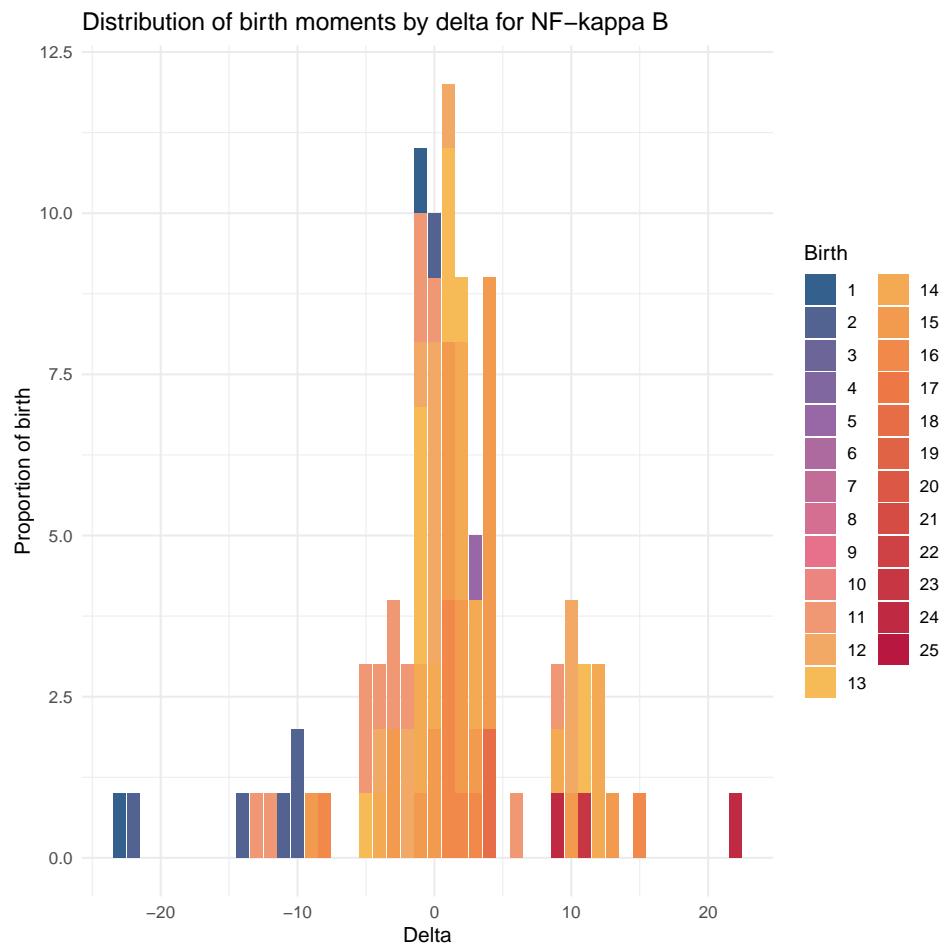
Distribution of birth moments by delta for Hippo



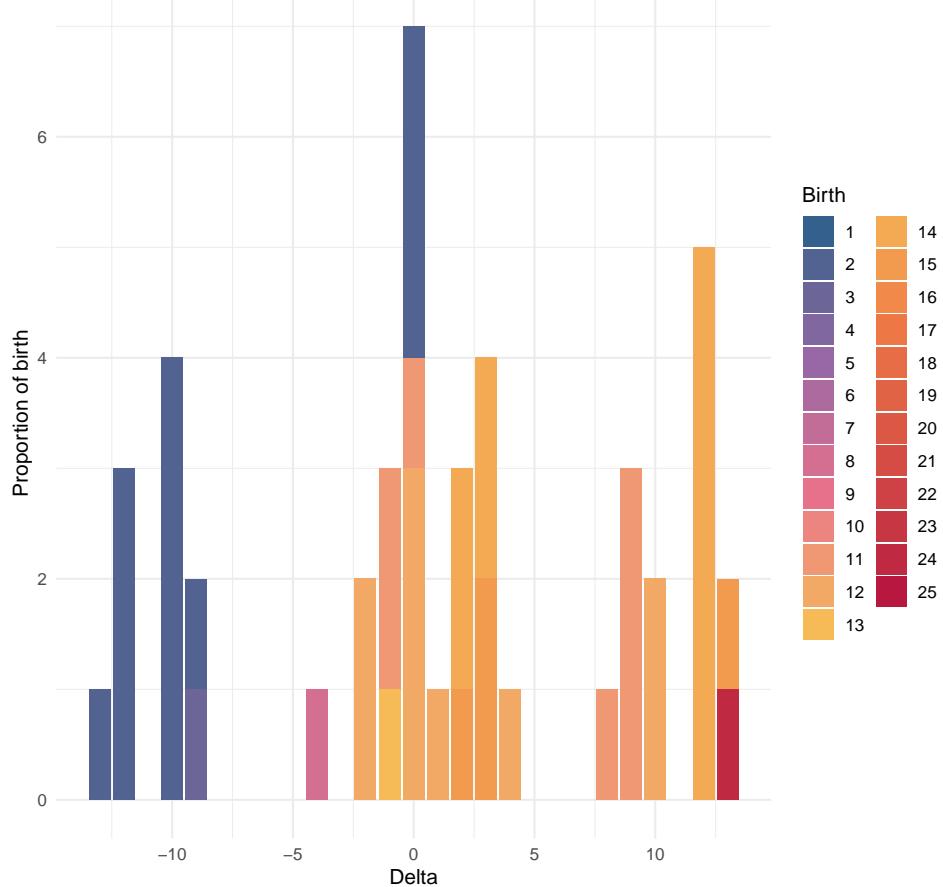
Distribution of birth moments by delta for JAK-STAT



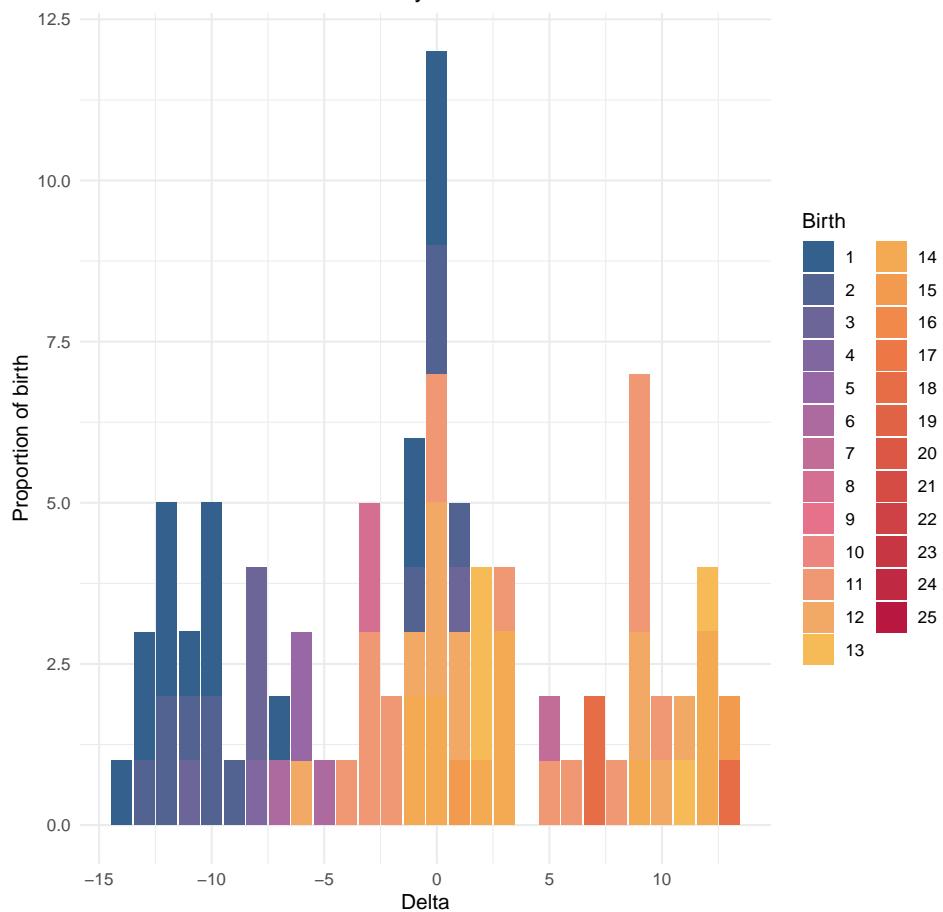




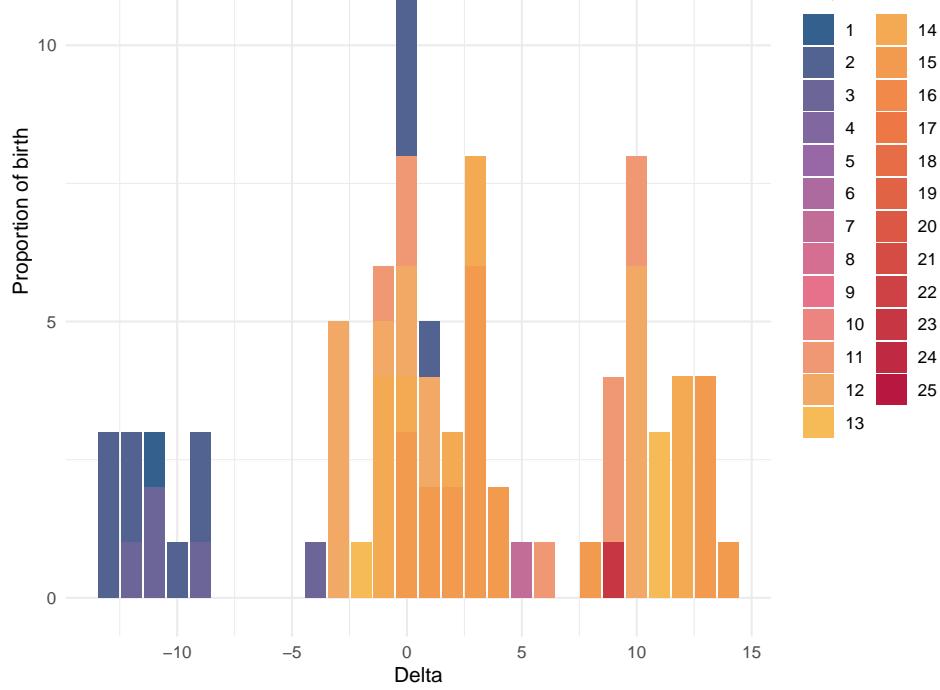
Distribution of birth moments by delta for Phospholipase D



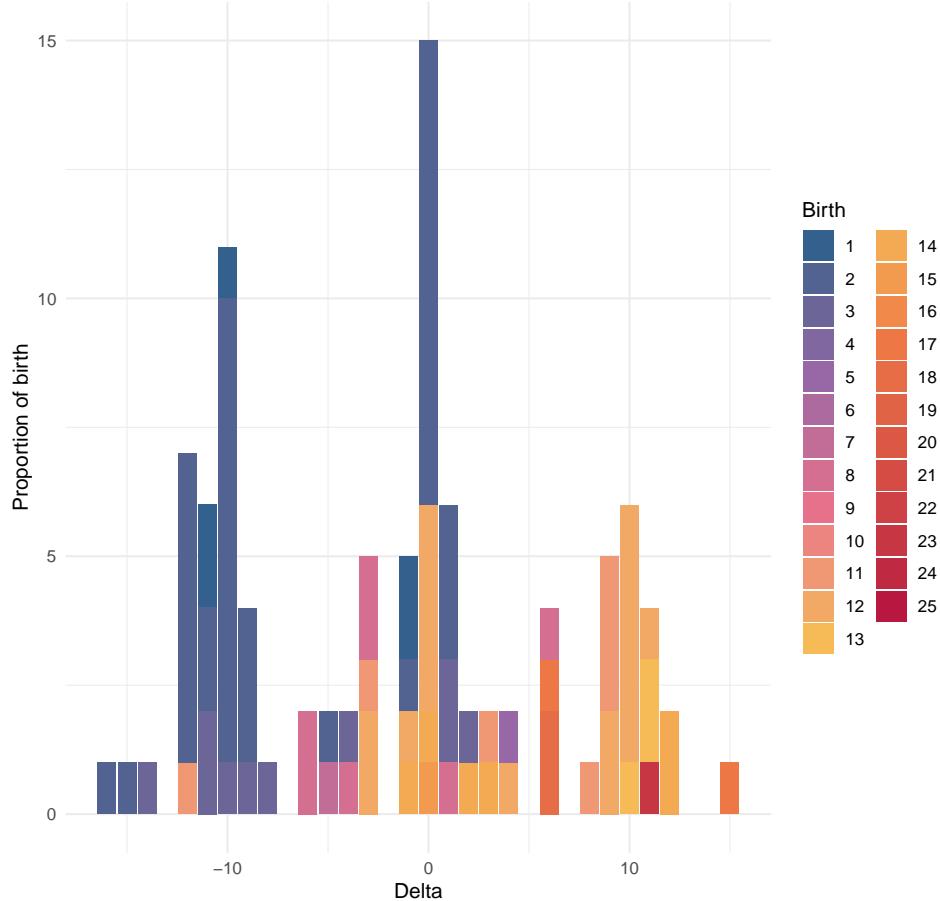
Distribution of birth moments by delta for PI3K-Akt



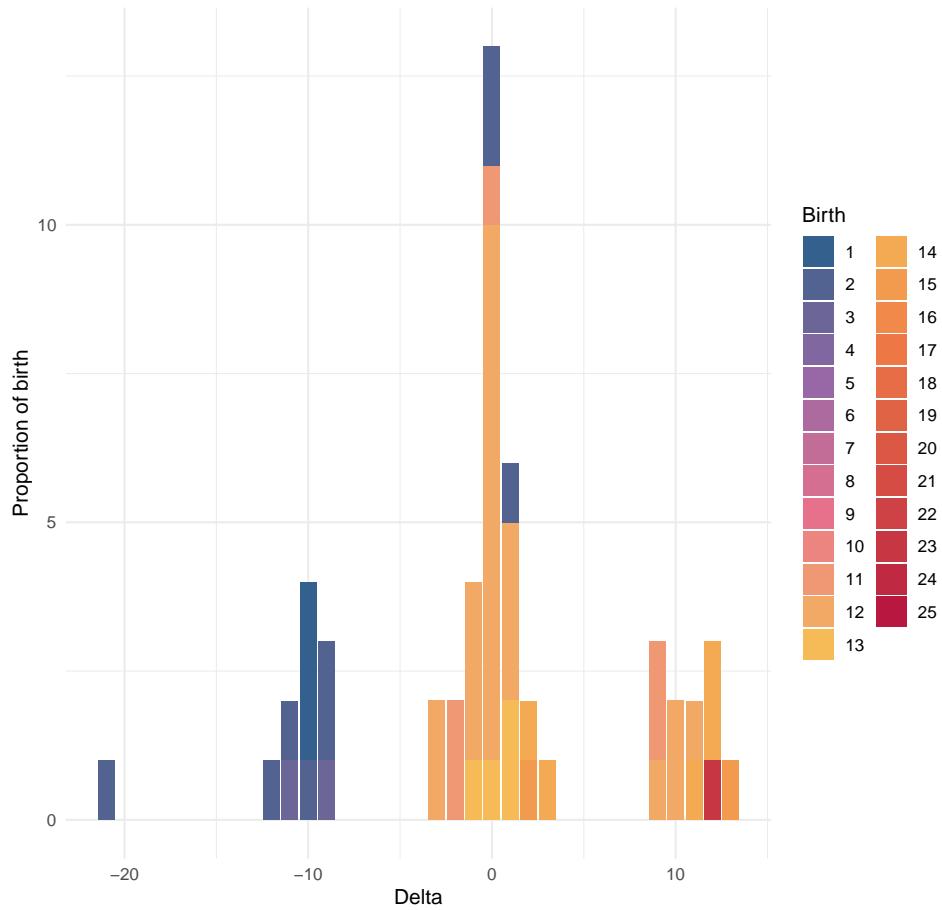
Distribution of birth moments by delta for Rap1



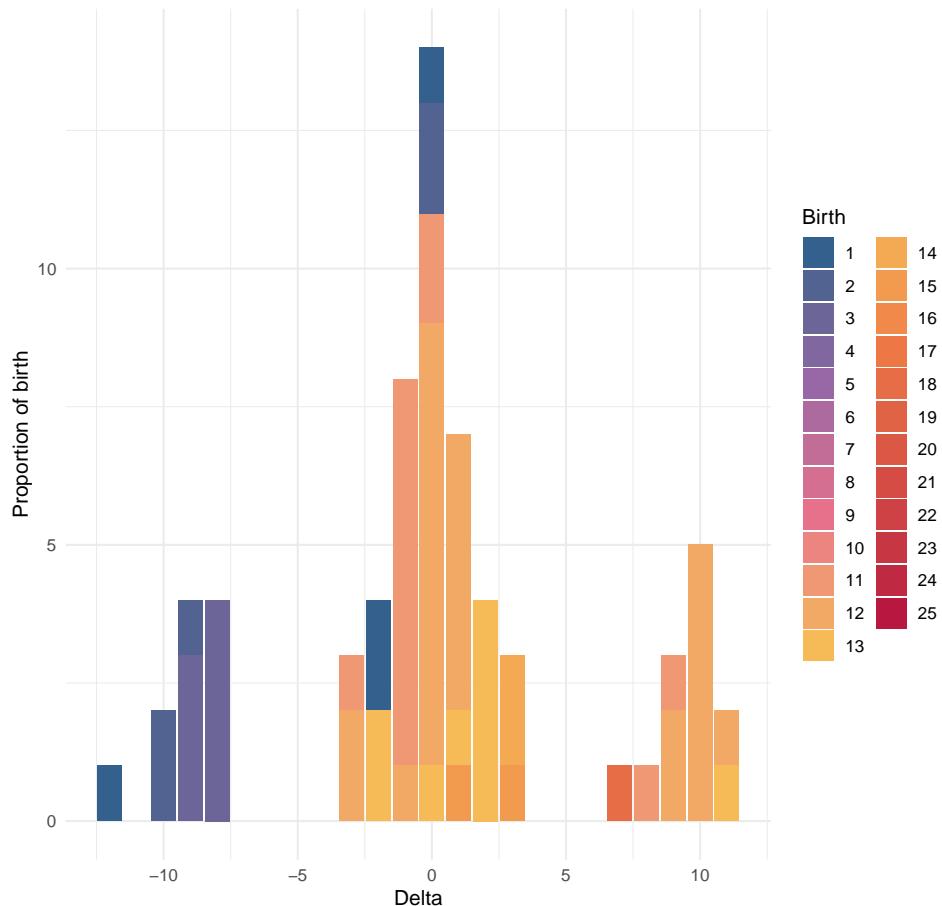
Distribution of birth moments by delta for Ras



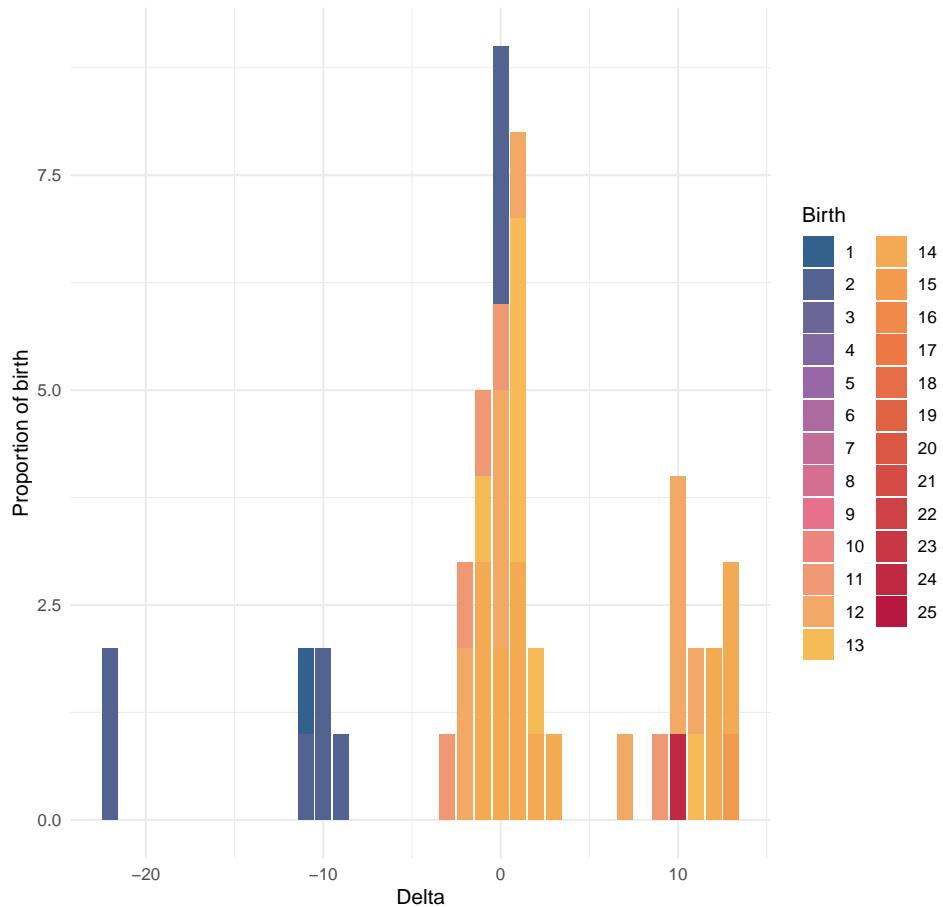
Distribution of birth moments by delta for Sphingolipid



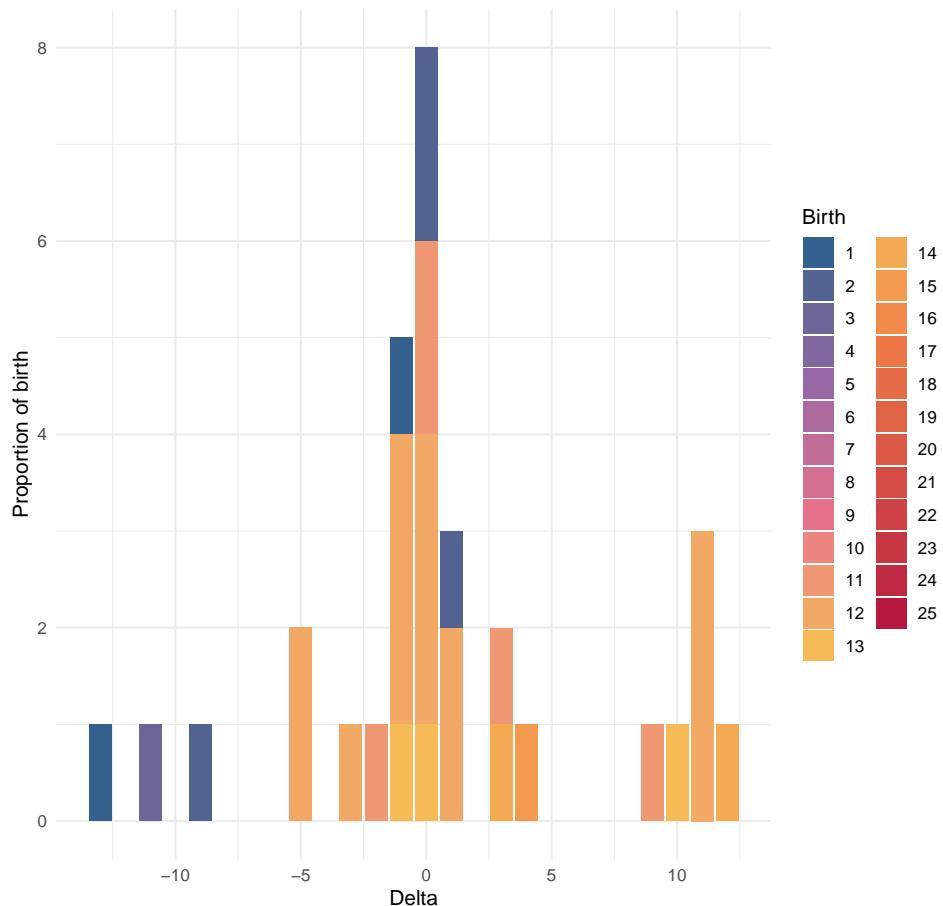
Distribution of birth moments by delta for TGF-beta



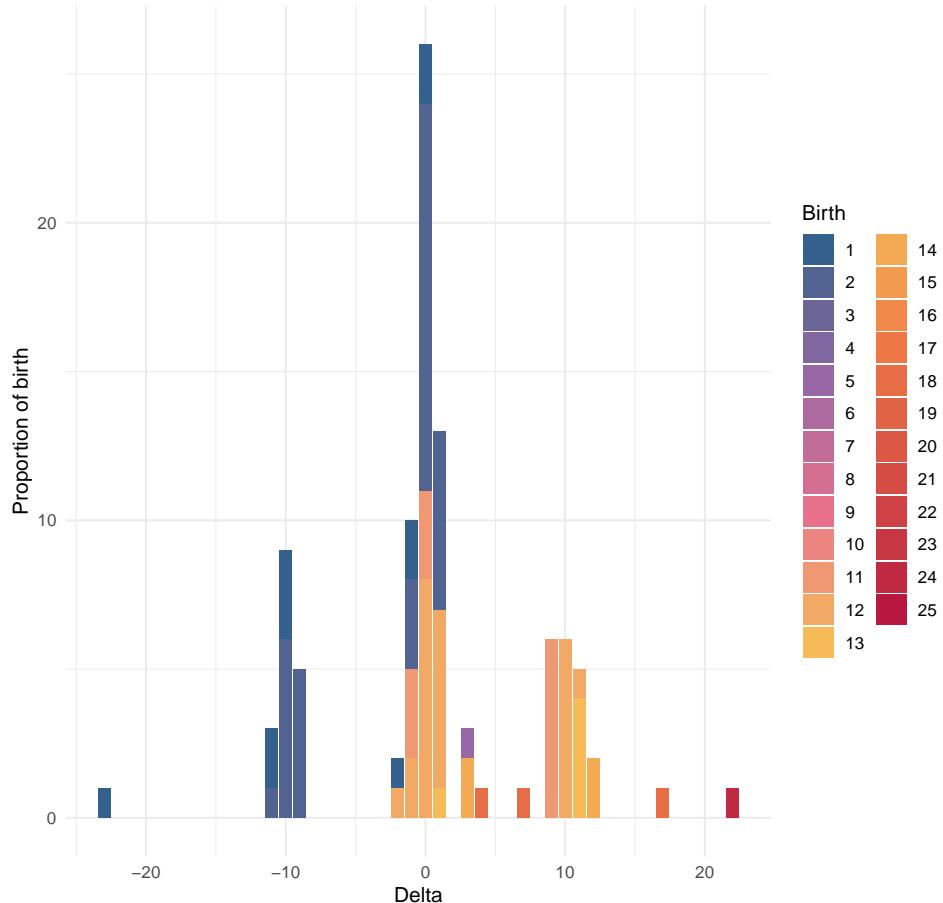
Distribution of birth moments by delta for TNF



Distribution of birth moments by delta for VEGF



Distribution of birth moments by delta for Wnt

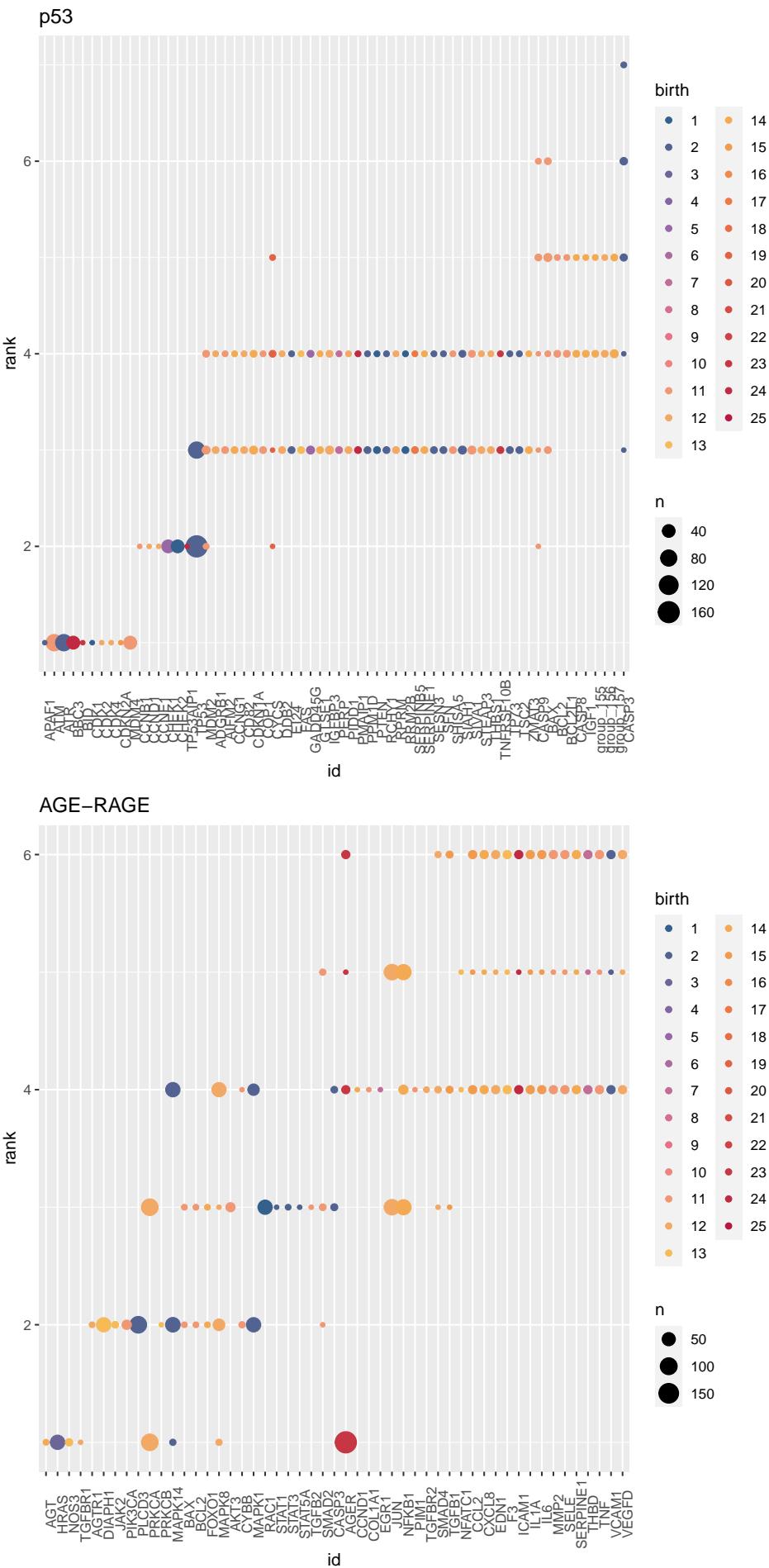


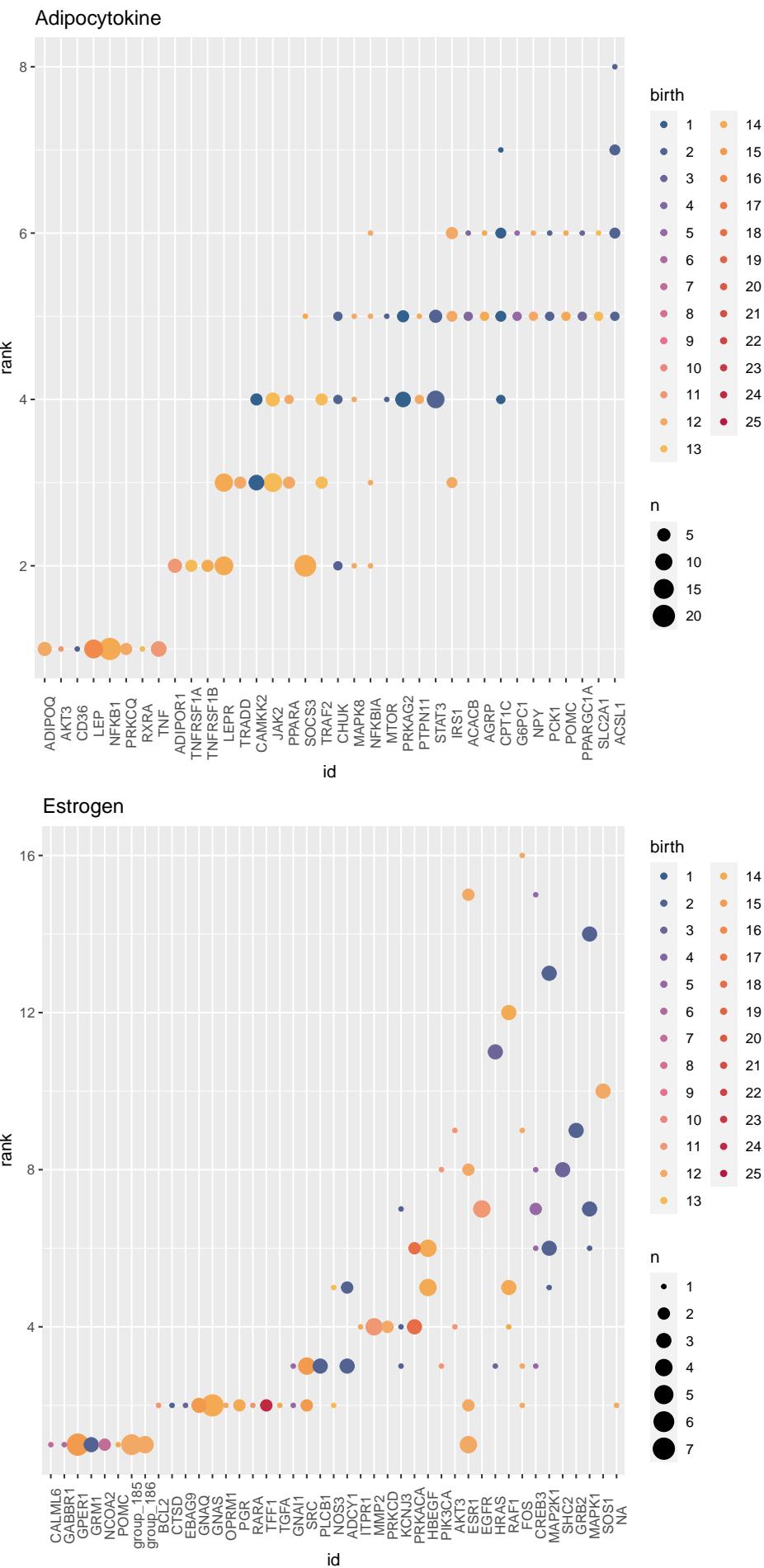
### 6.3 Article 1 - Suppl. Data 2

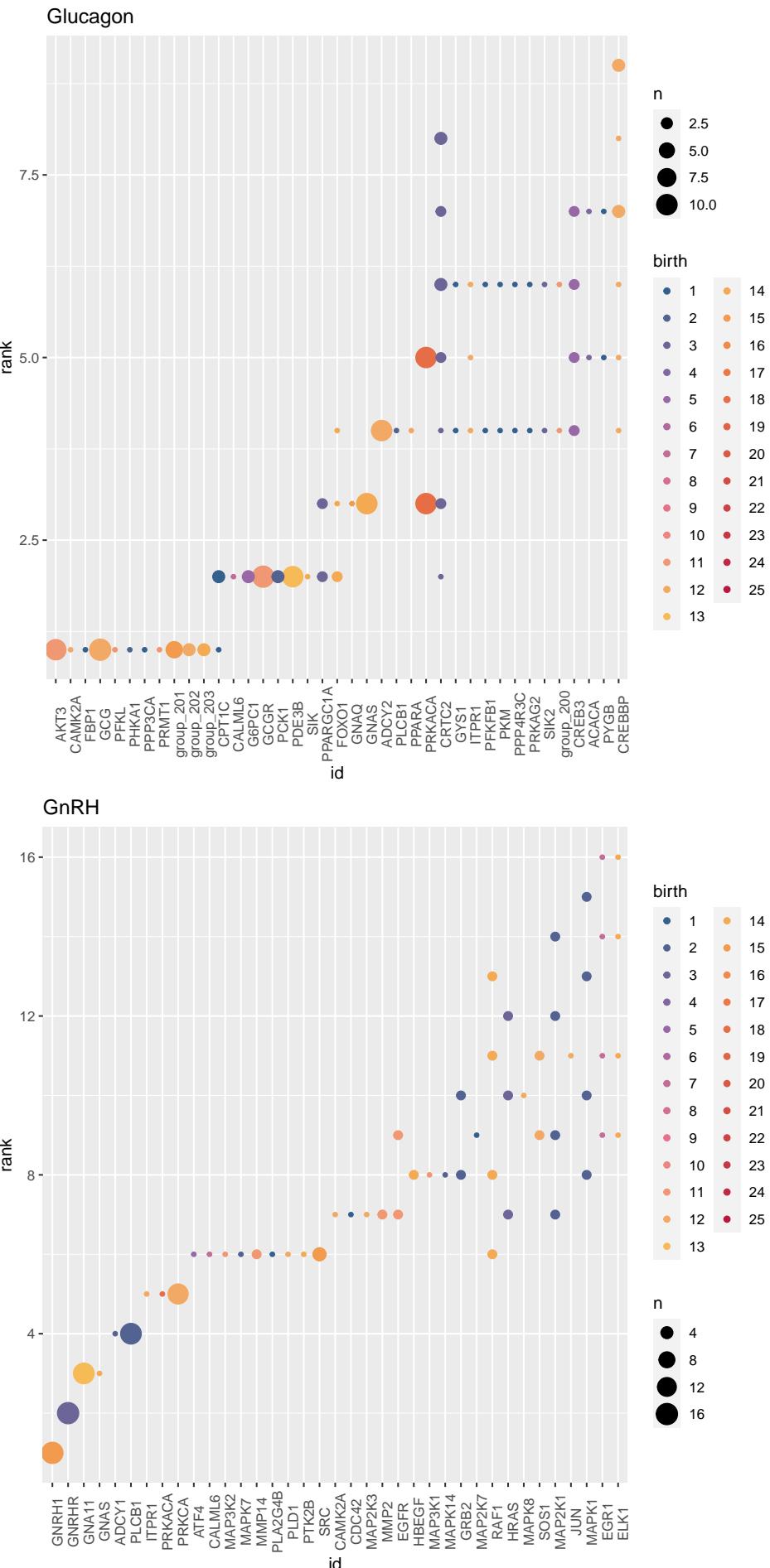
#### Suppl. Data 2 - Distribution of genes by position/rank in the pathway and node of birth

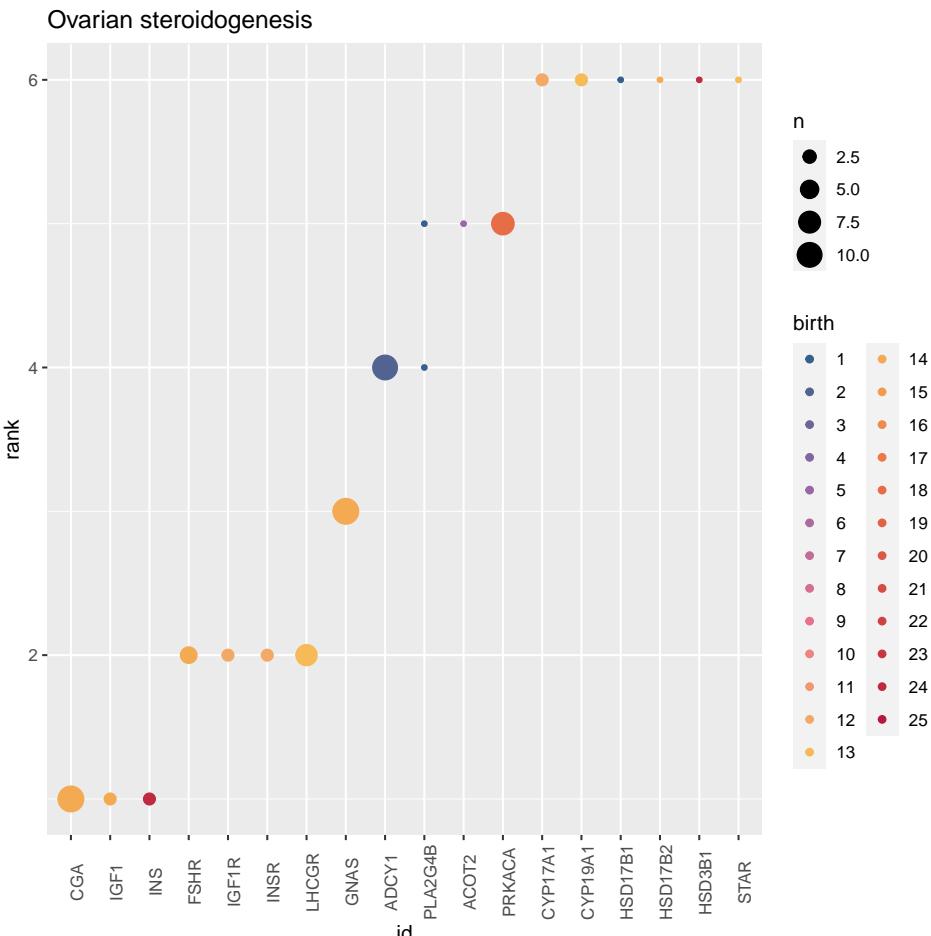
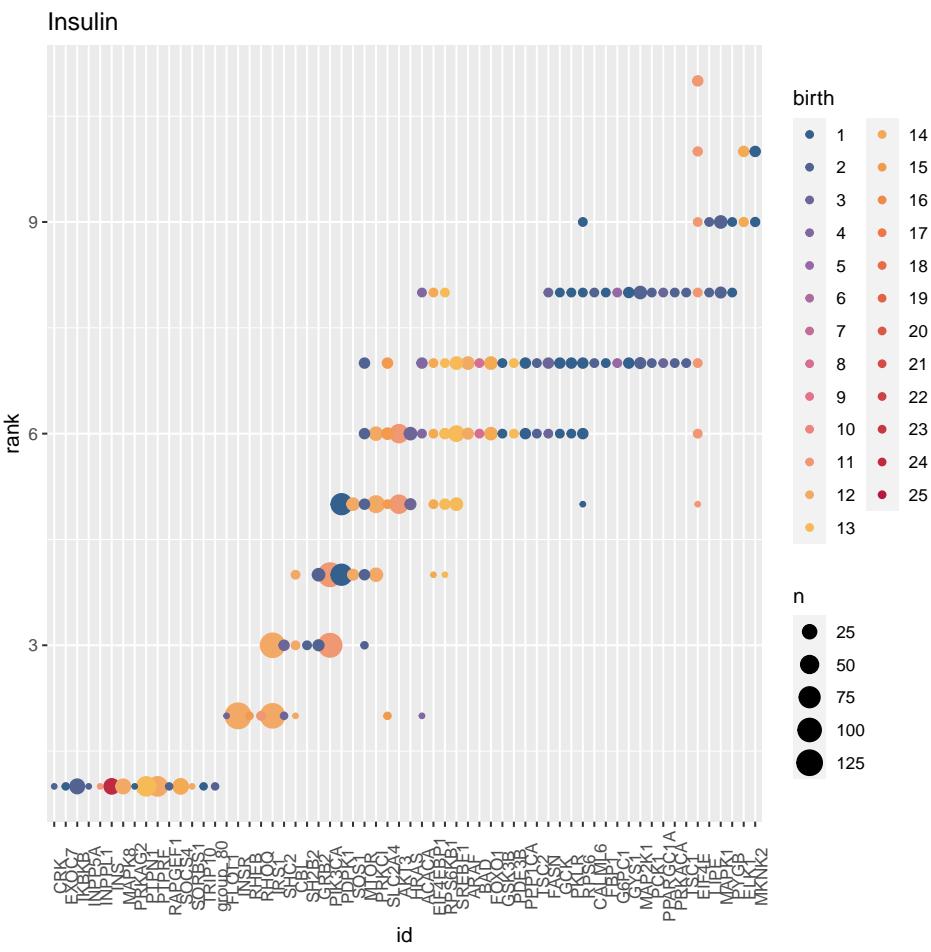
Legend : Each graph represents one of the 47 pathways. Abscissa : the different proteins involved in the pathway; ordinate : position/rank of the protein within the pathway. Each protein is colored depending on the node of birth of its corresponding gene. Proteins are represented by a dot, are characterized by their position(s) they occupy within the pathway. The size of the dots is proportional to the number of times they are in these positions.

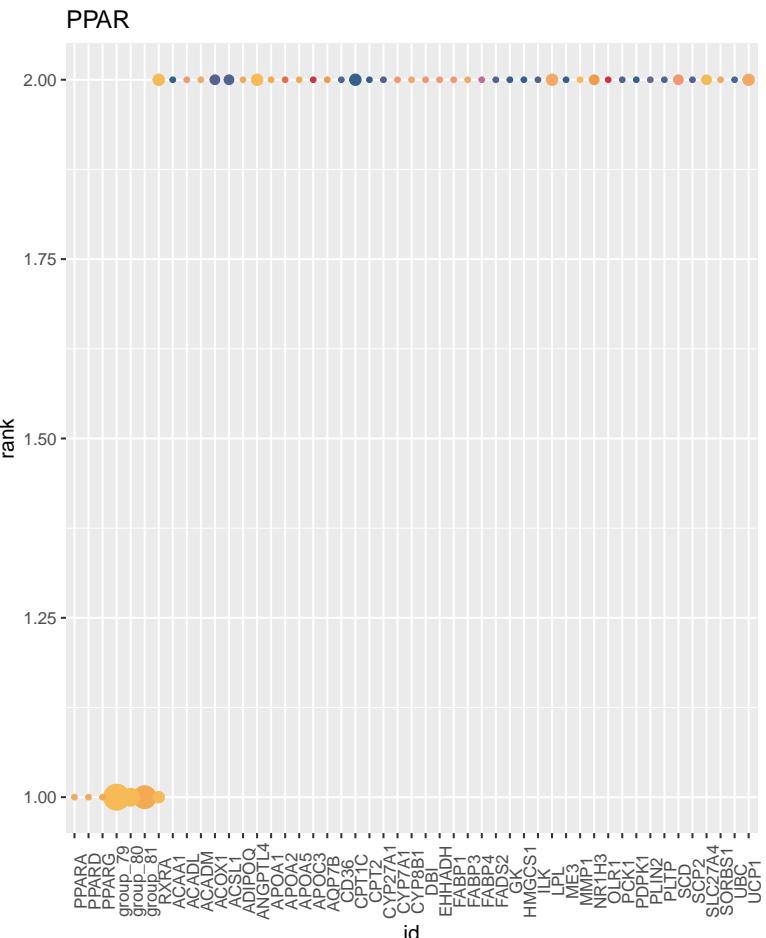
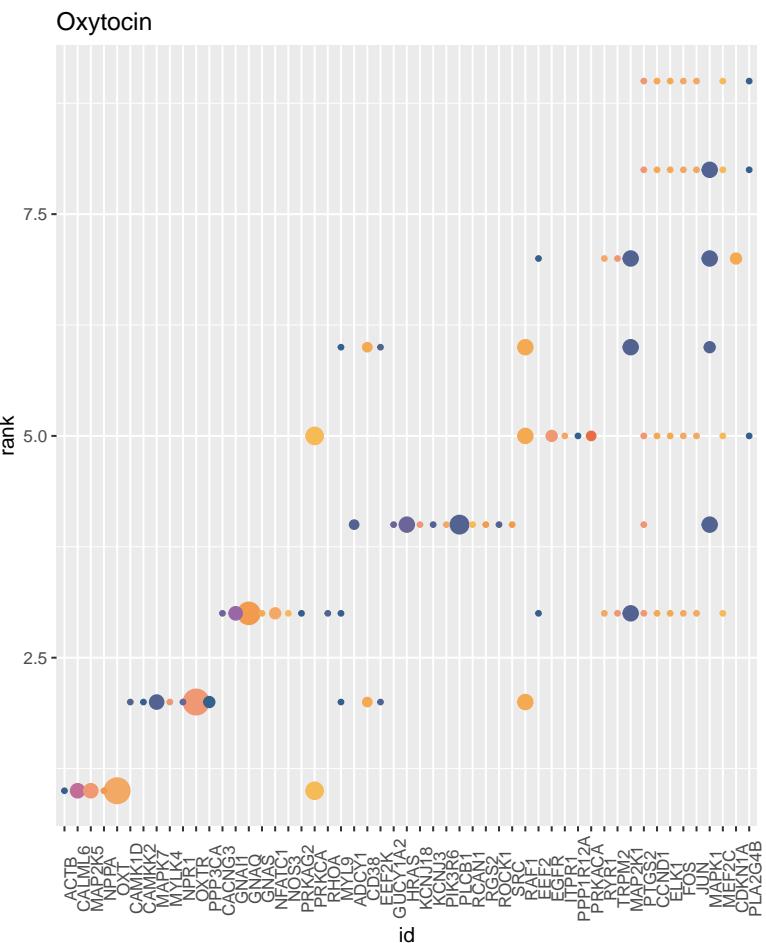
The pathways are in the following order : p53, AGE-RAGE, Adipocytokine, Estrogen, Glucagon, GnRH, Insulin, Ovarian steroidogenesis, Oxytocin, PPAR, Prolactin, Relaxin, Thyroid hormone, B cell receptor, C-type lectin receptor, Chemokine, FC epsilon RI, IL-17, NOD-like receptor, RIG-I-like receptor, T cell receptor, Toll-like receptor, Neurotrophin, AMPK, Apelin, Calcium, cAMP, cGMP-PKG, ErbB, FoxO, Hedgehog, HIF-1, Hippo, JAK-STAT, MAPK, mTOR, NF-Kappa B, Notch, Phospholipase D, PI3K-Akt, Rap1, Ras, Sphingolipid, TGF-Beta, TNF , VEGF, Wnt.

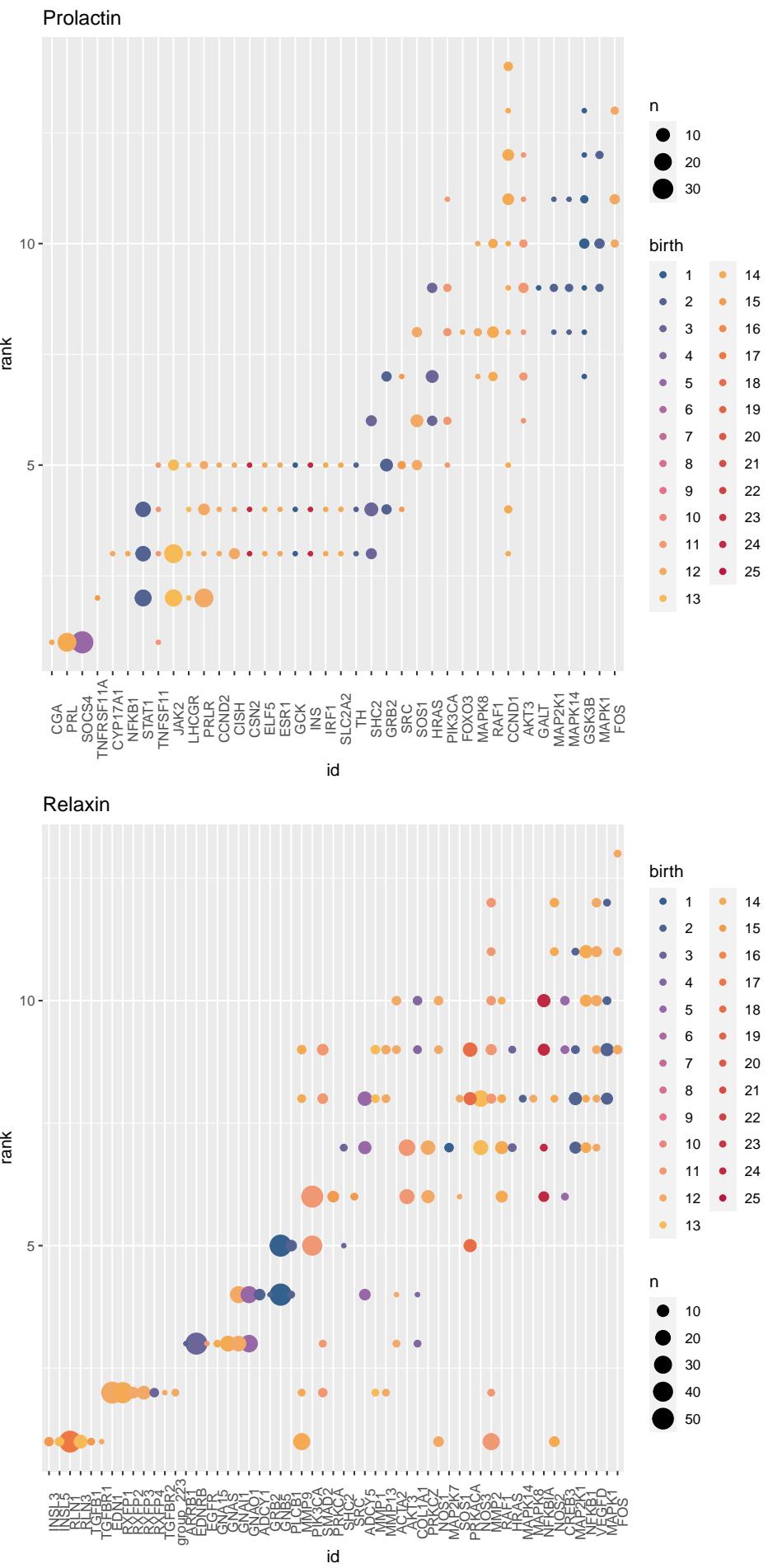


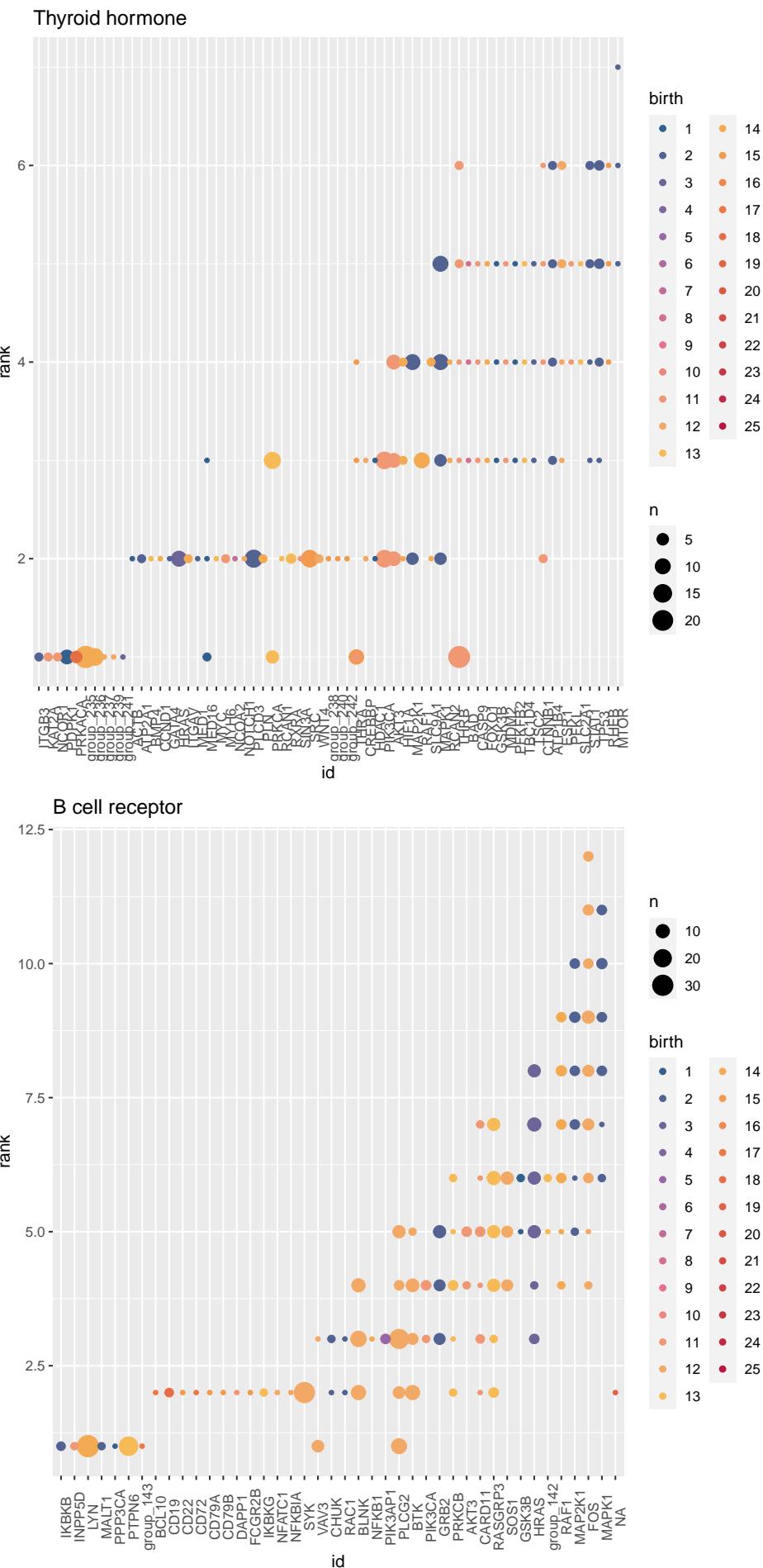




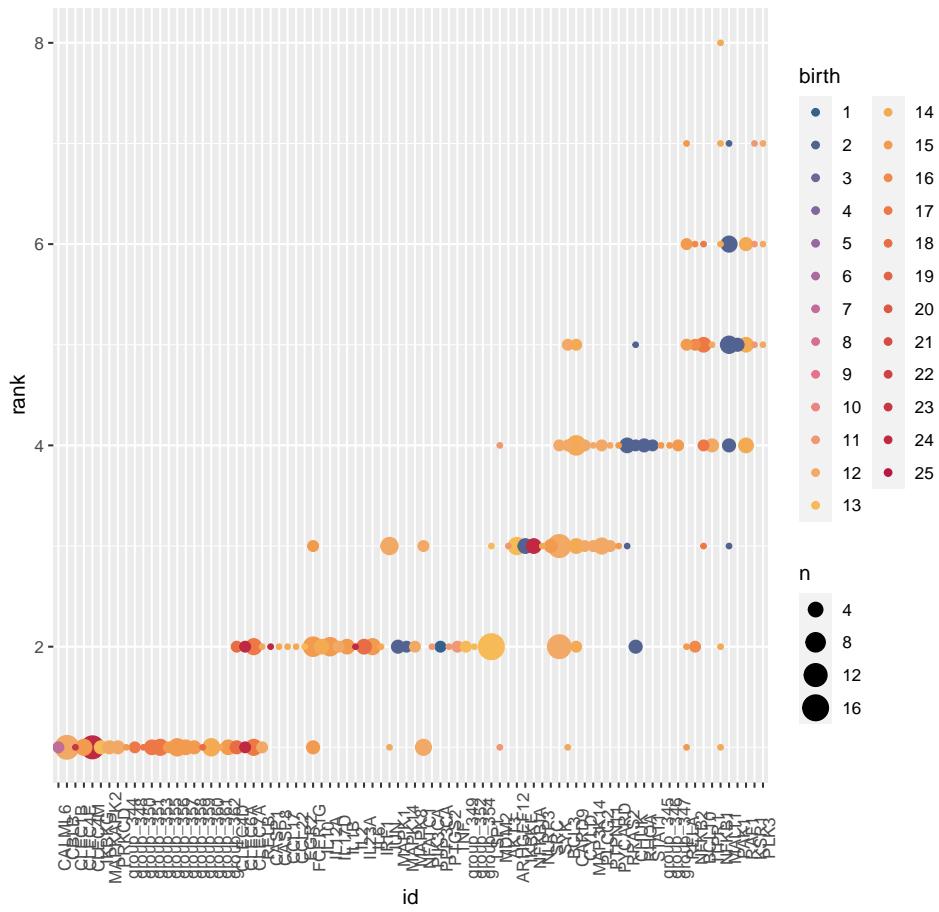




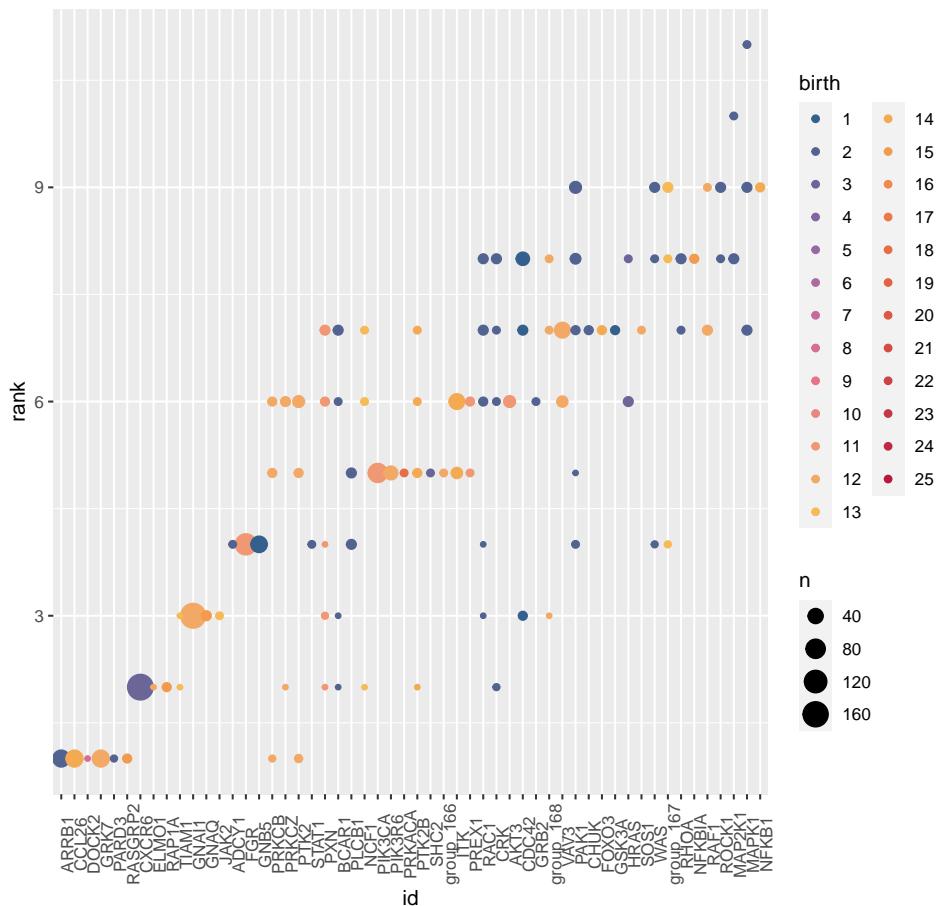


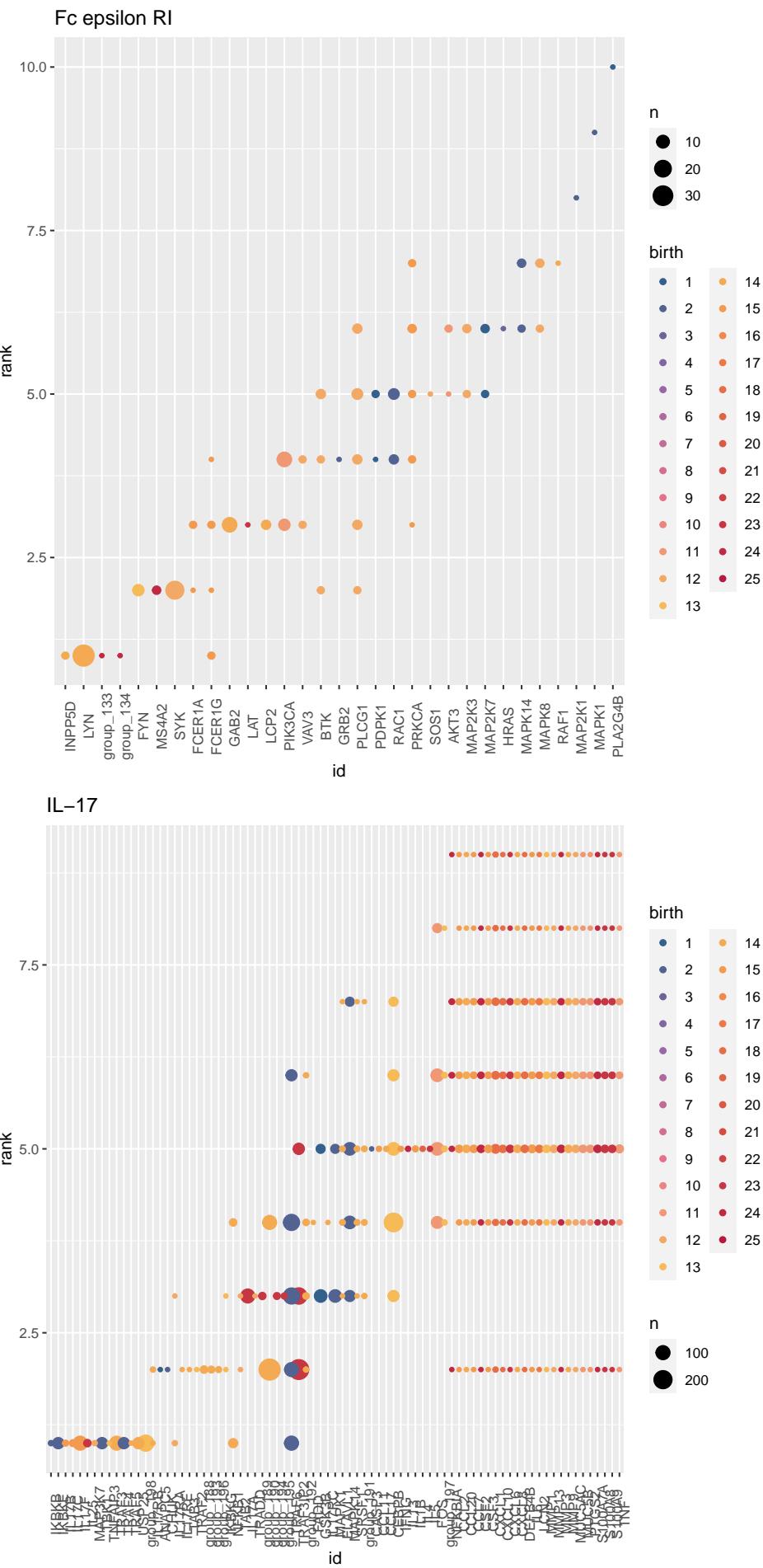


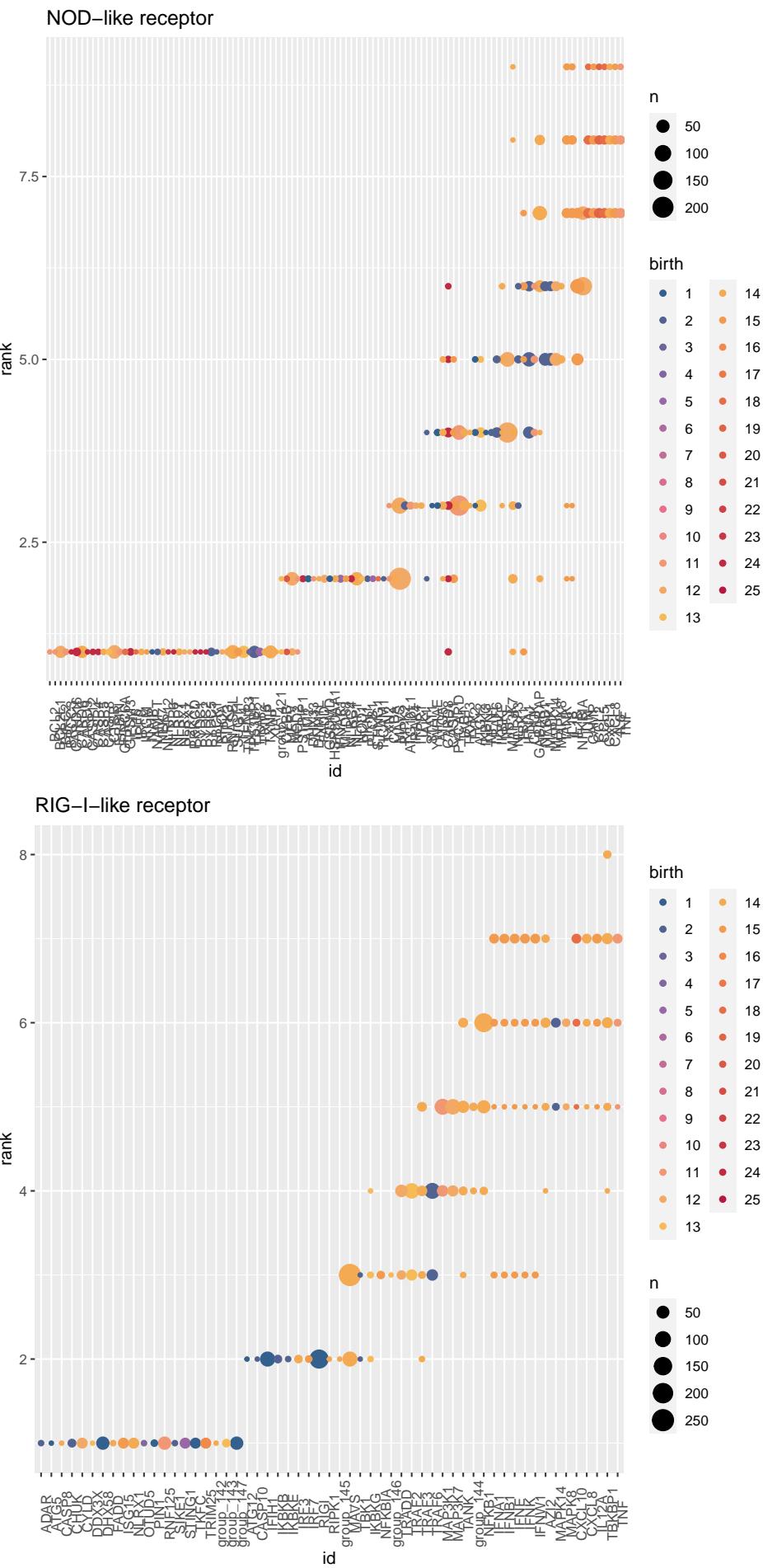
## C-type lectin receptor

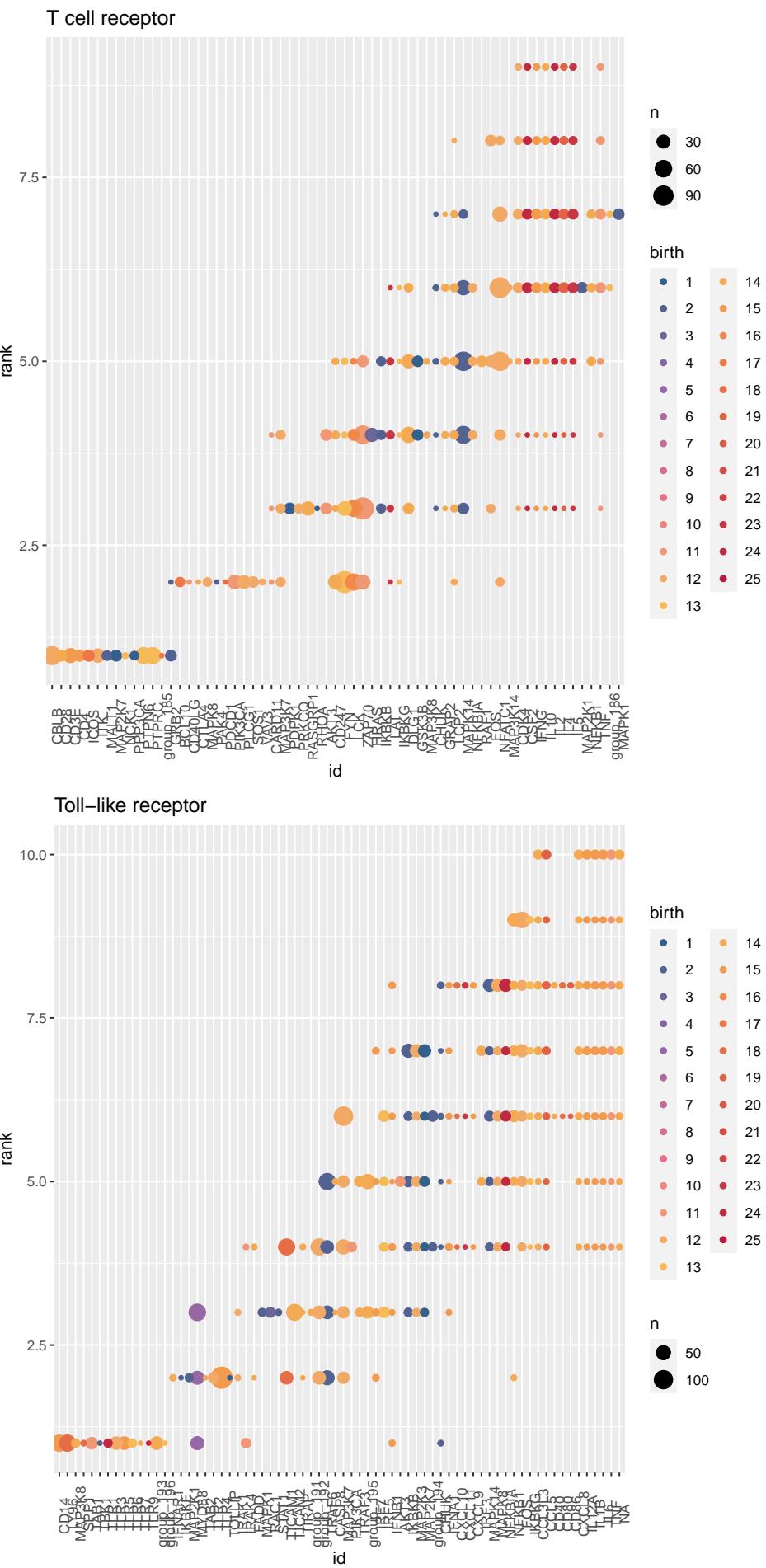


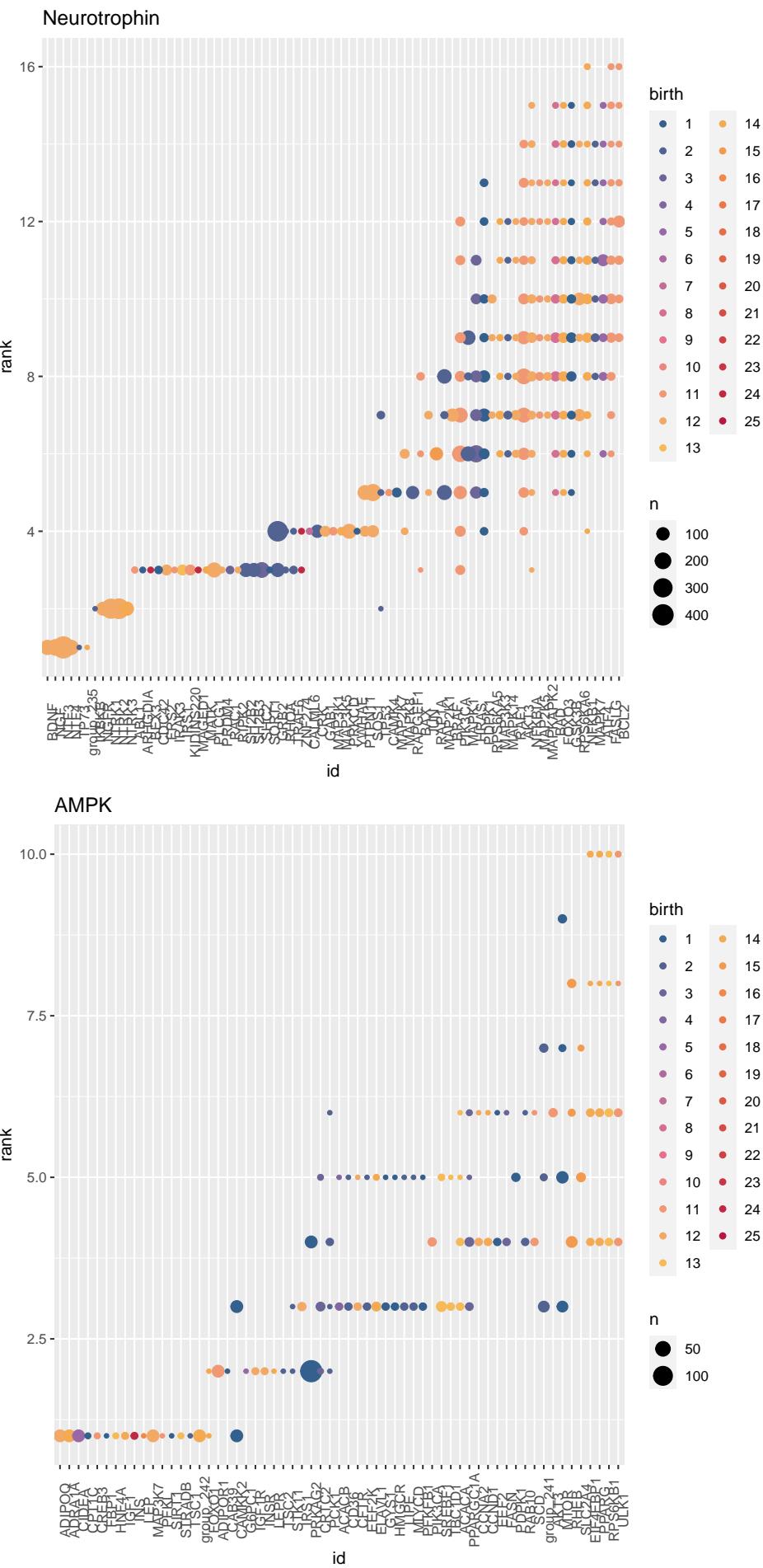
## Chemokine



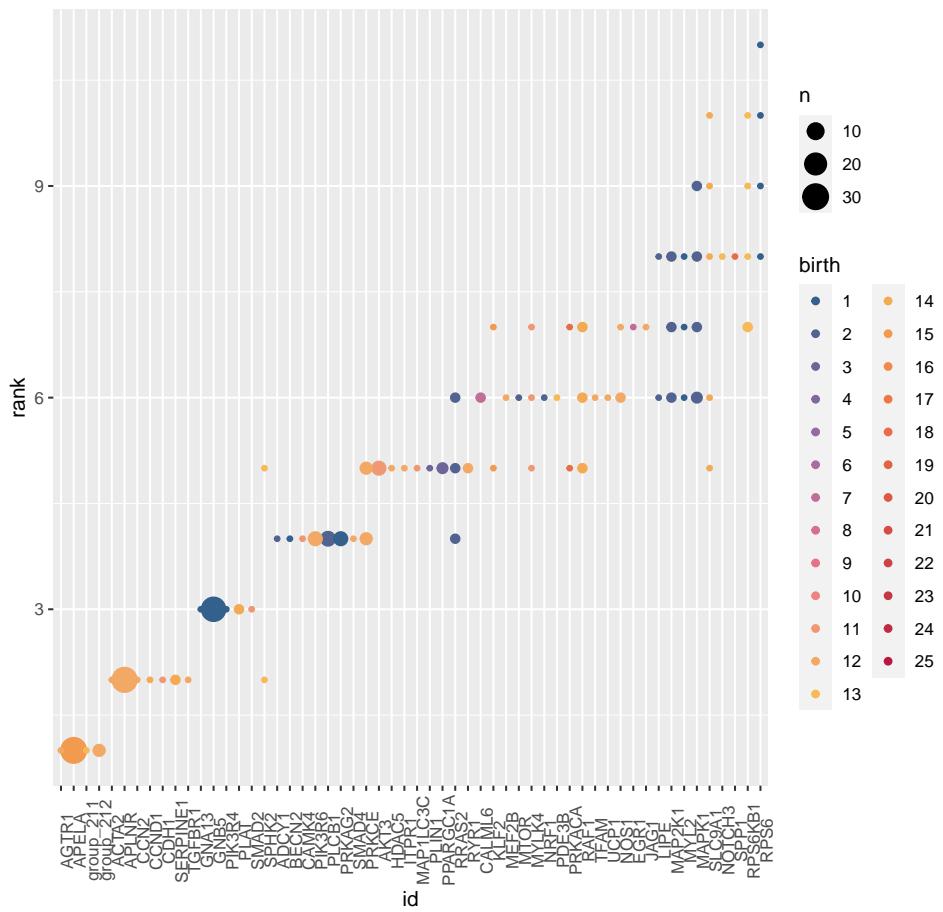




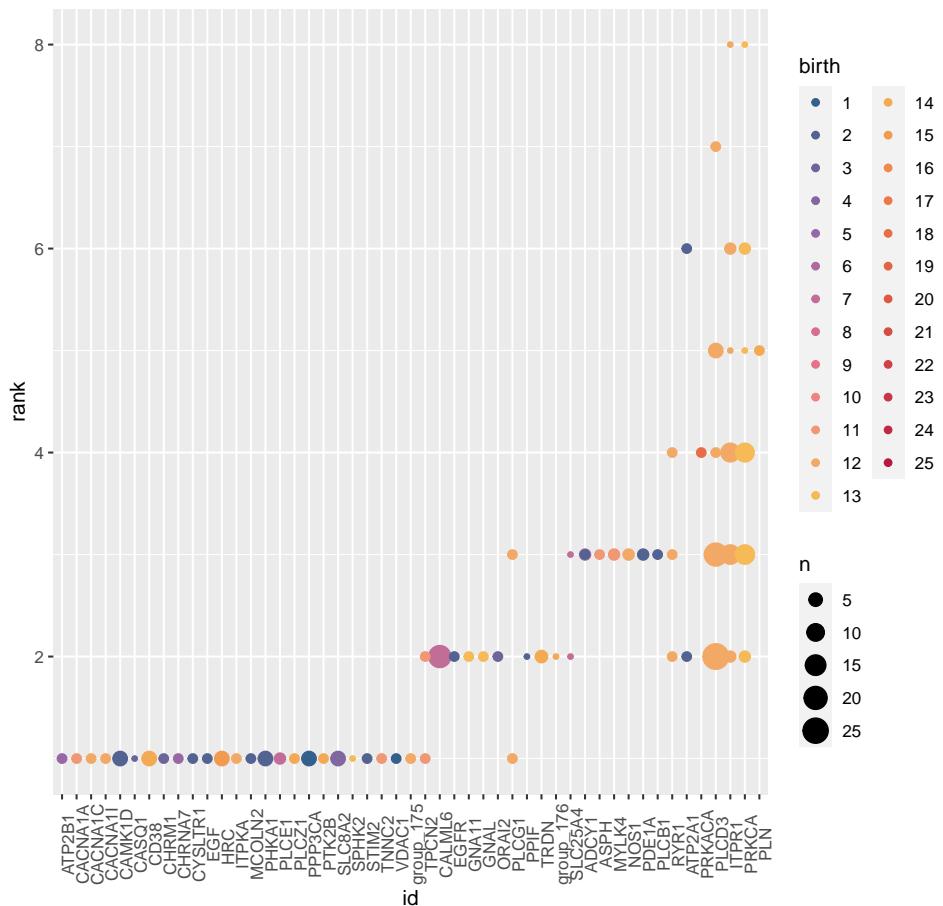


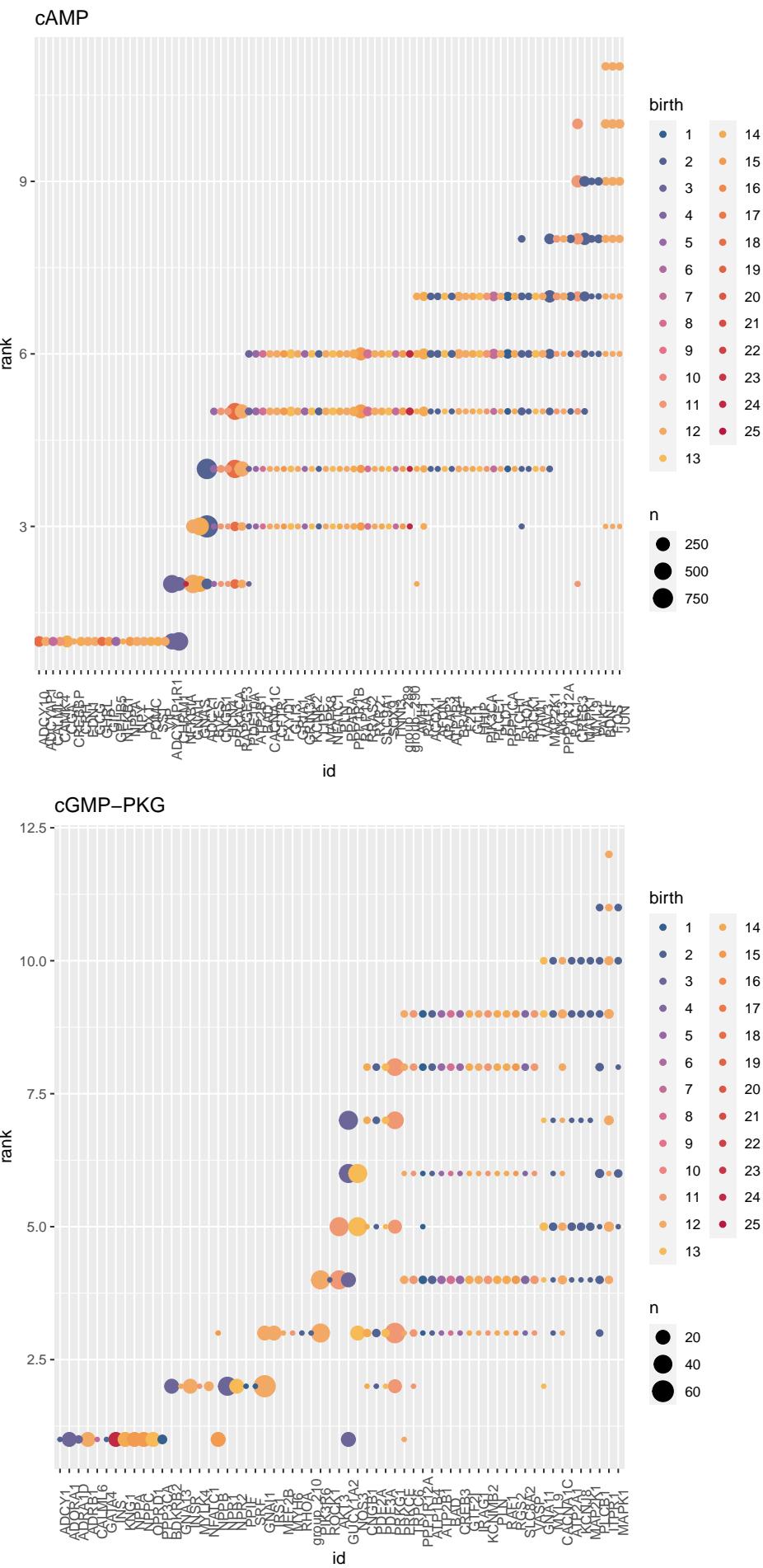


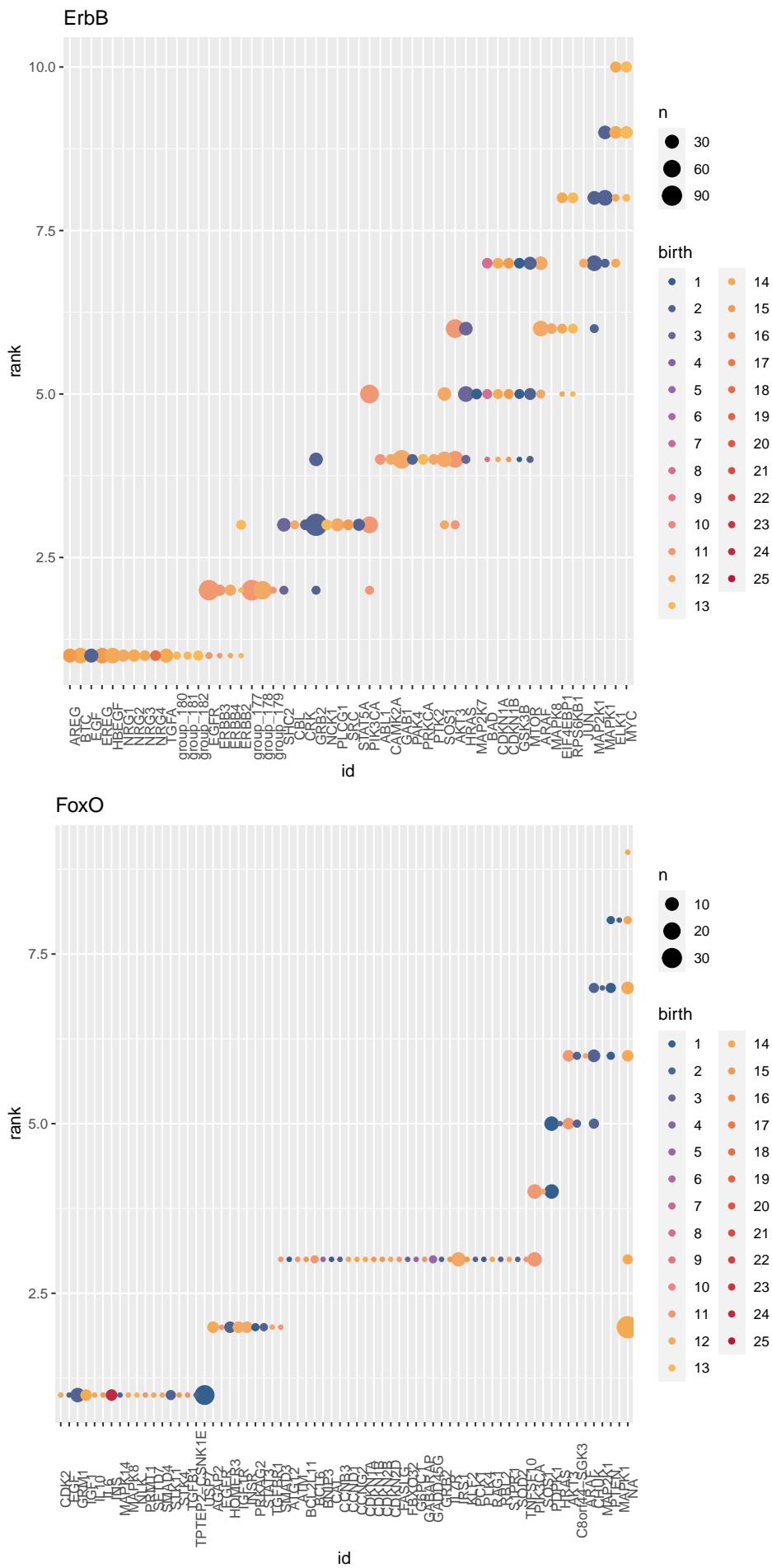
## Apelin

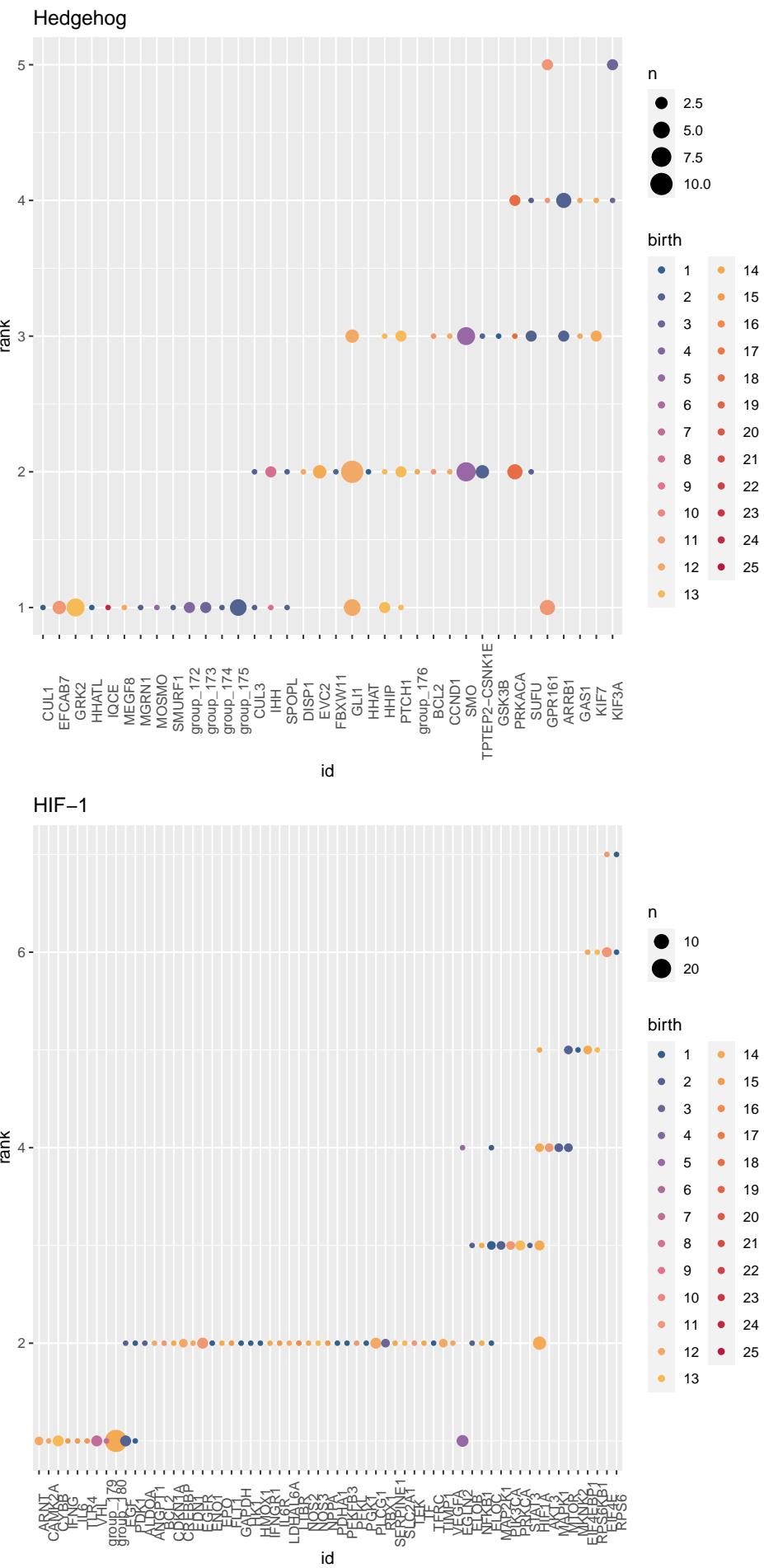


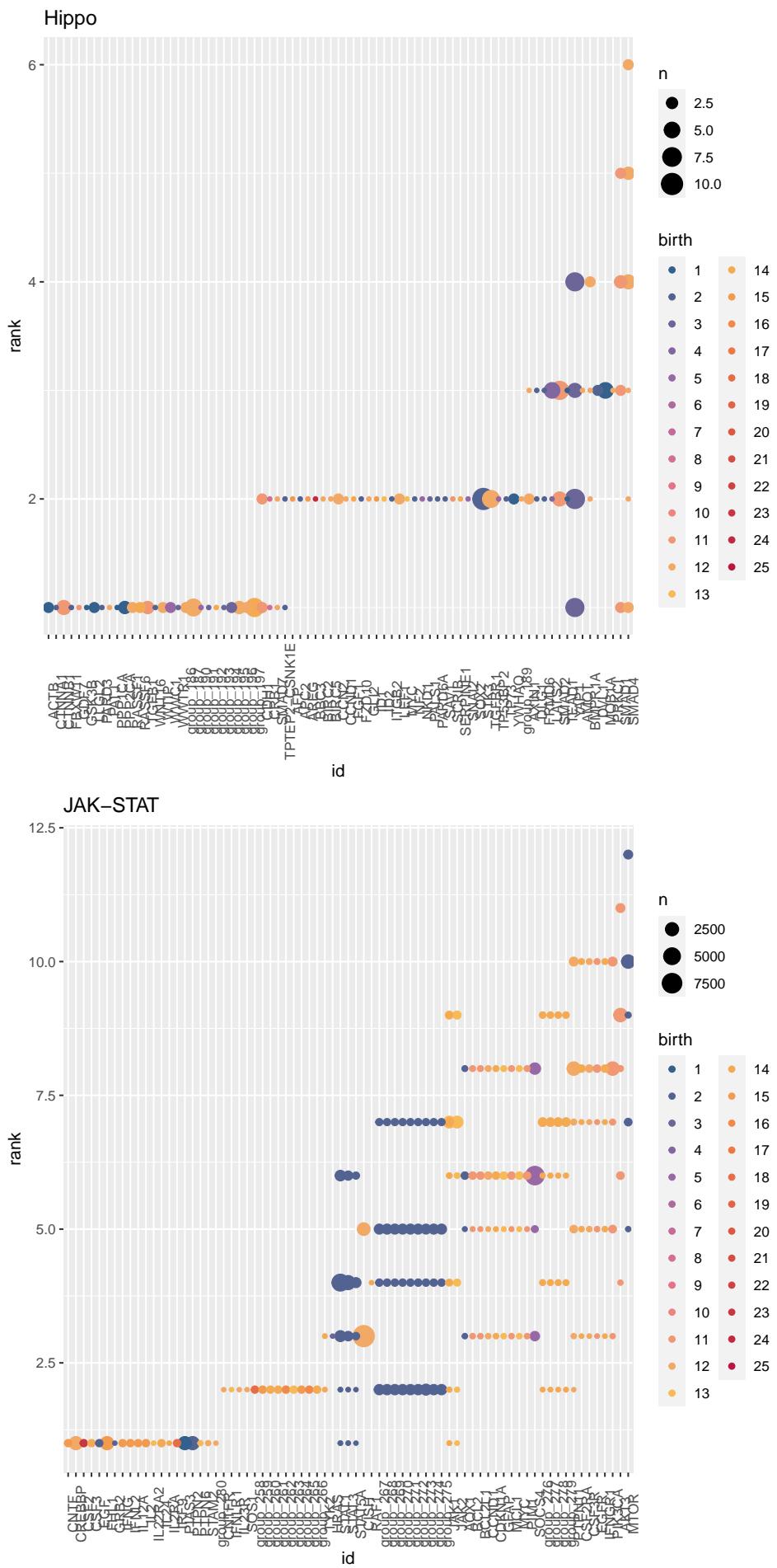
## Calcium



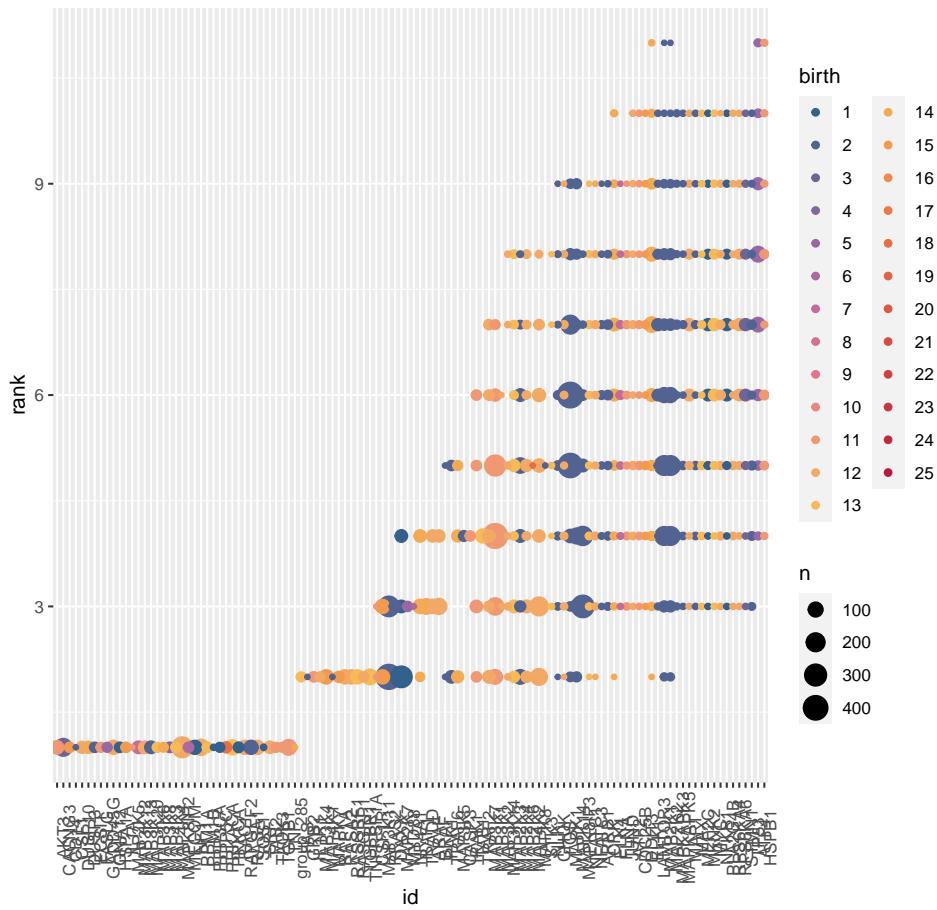




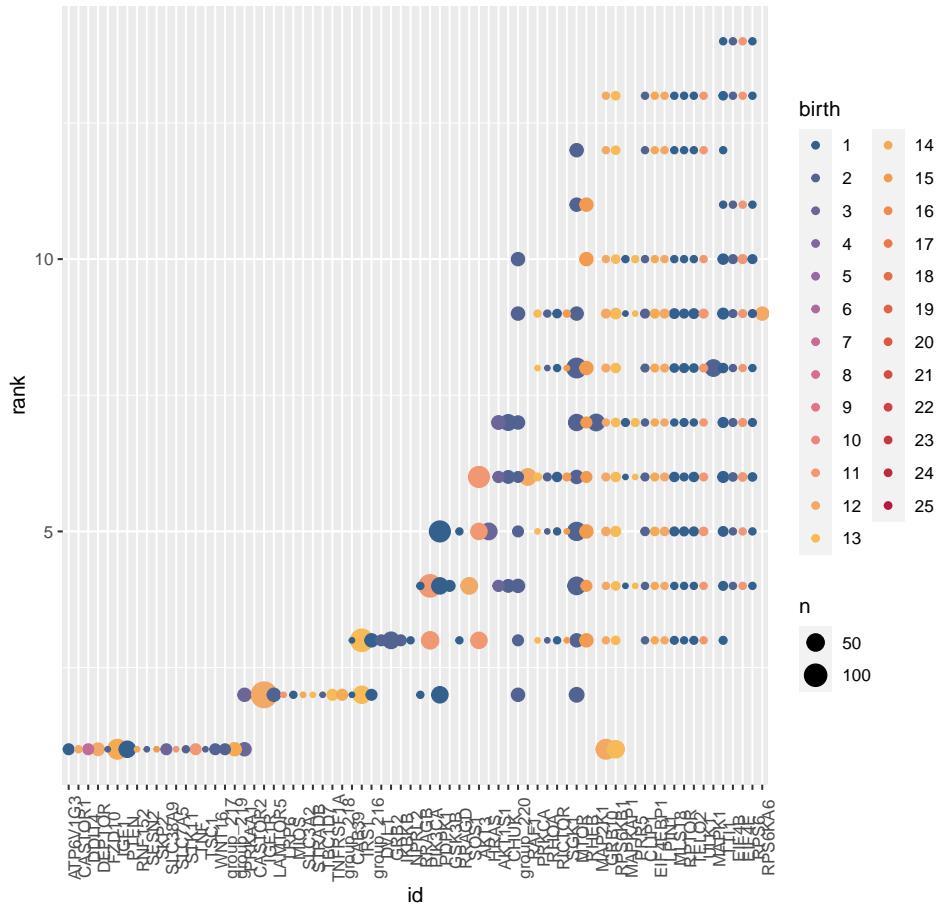




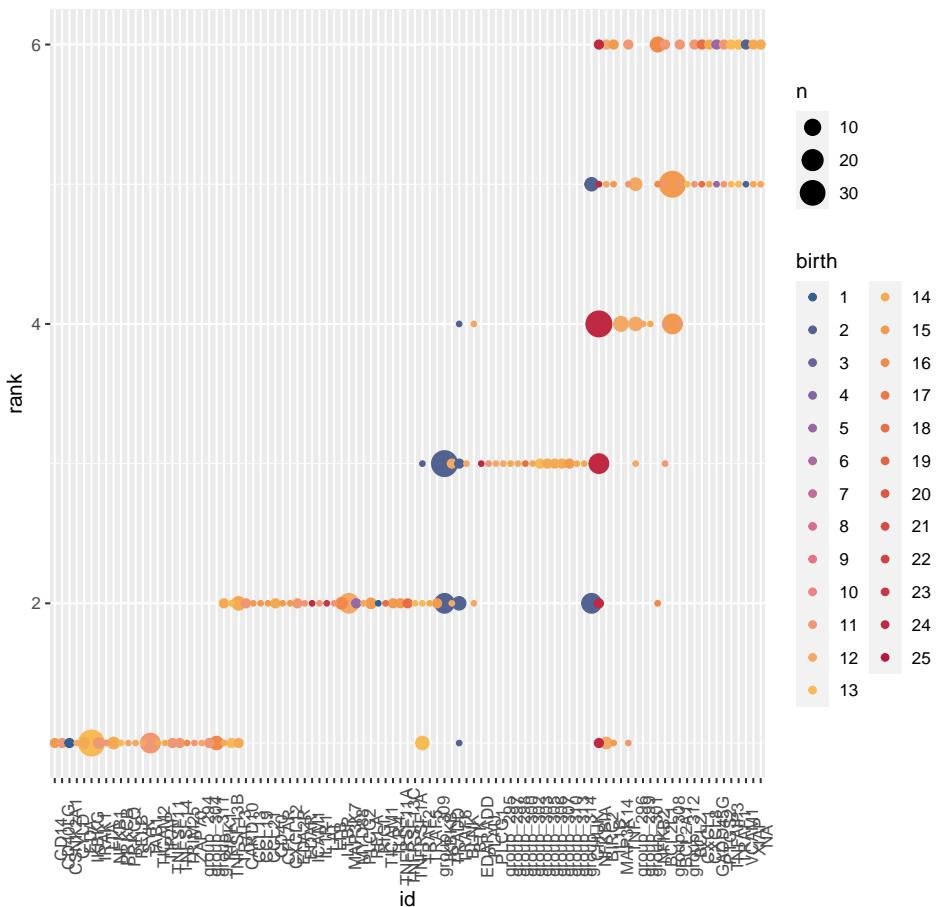
## MAPK



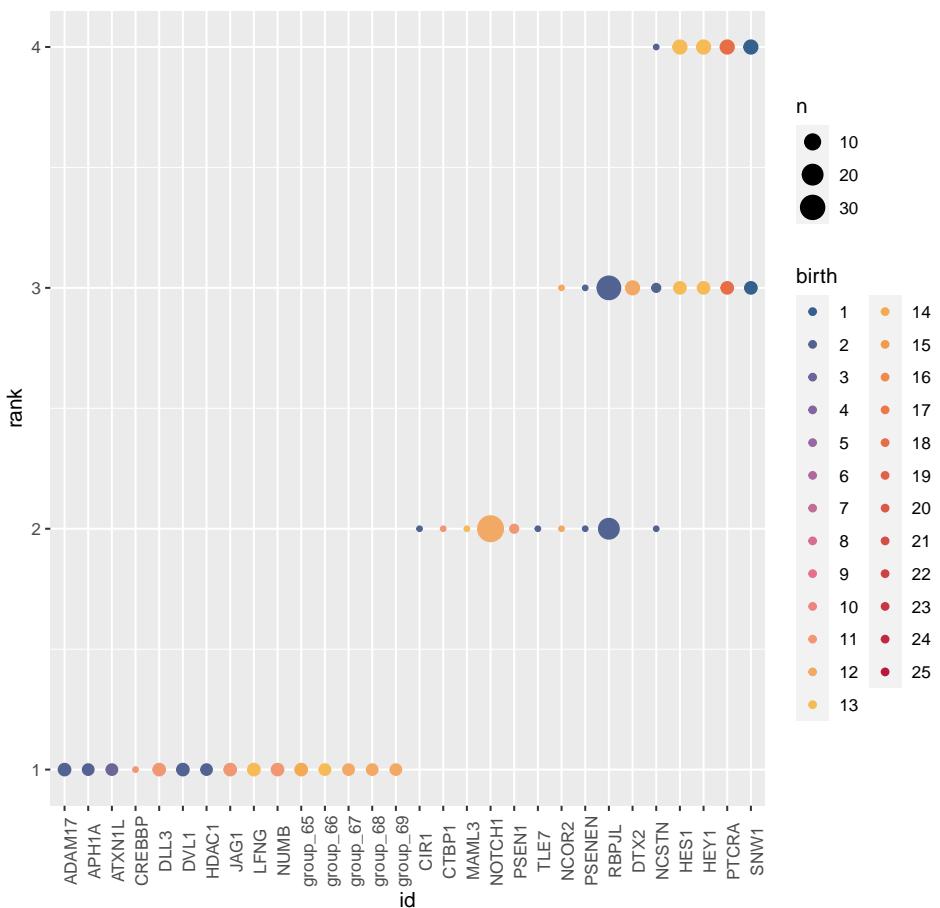
## mTOR



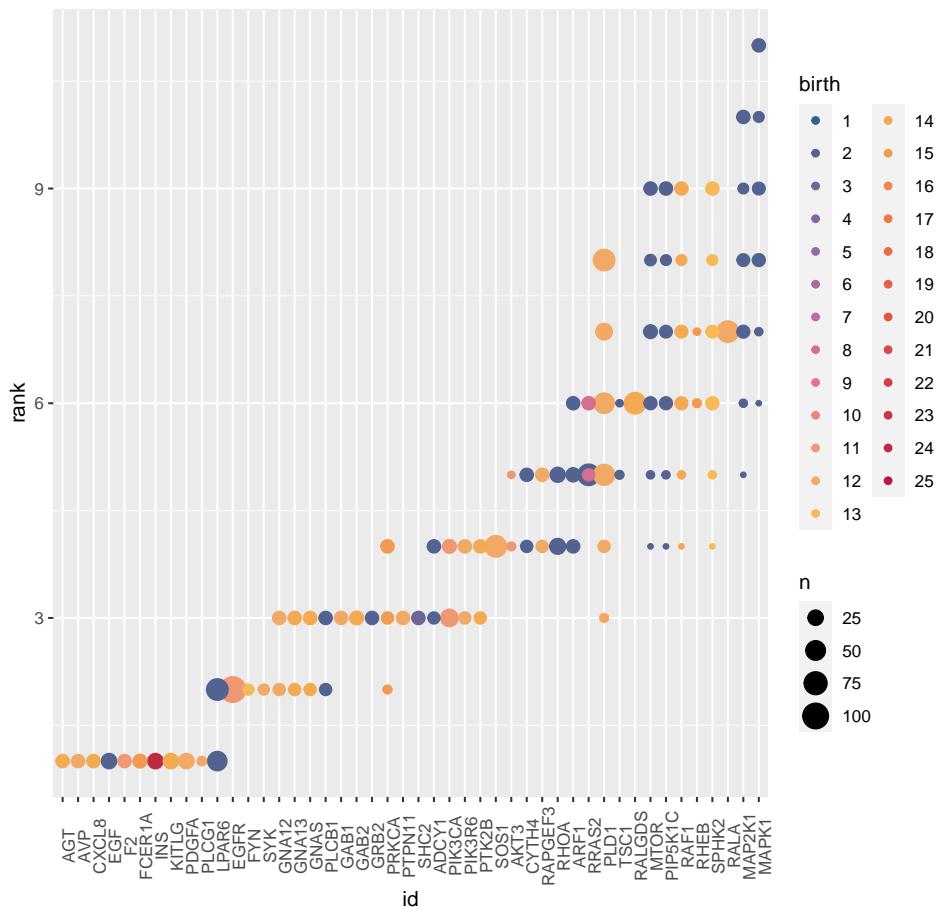
## NF-kappa B



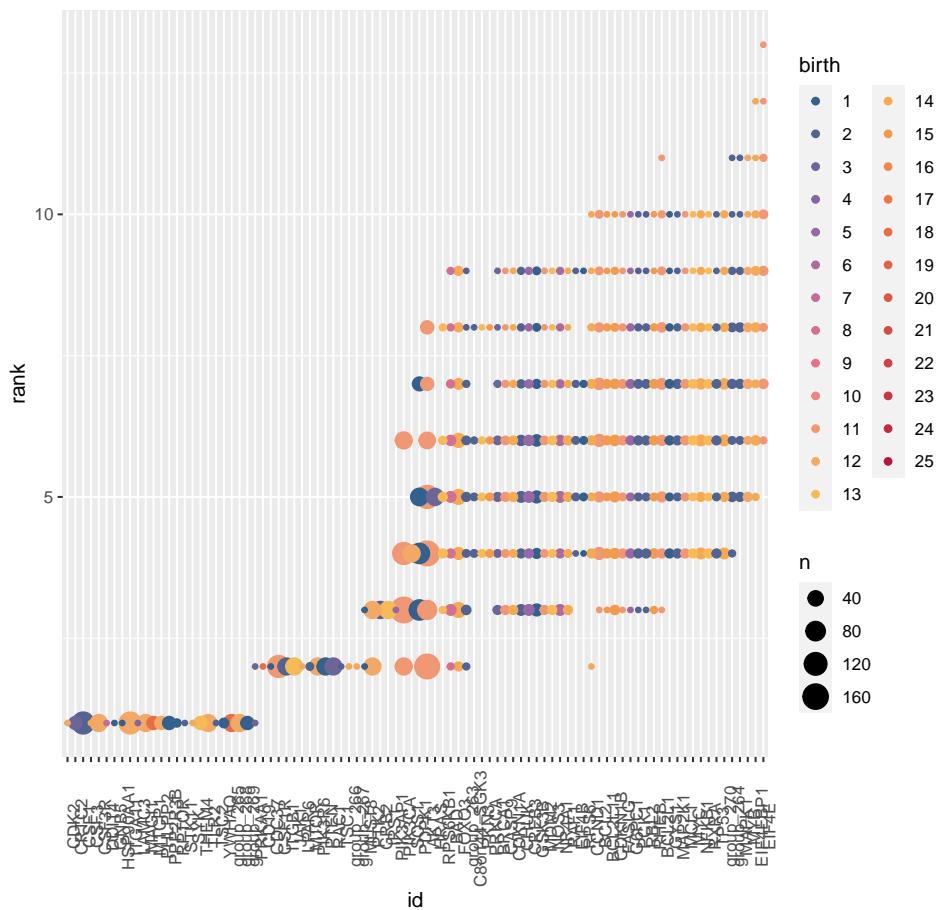
## Notch

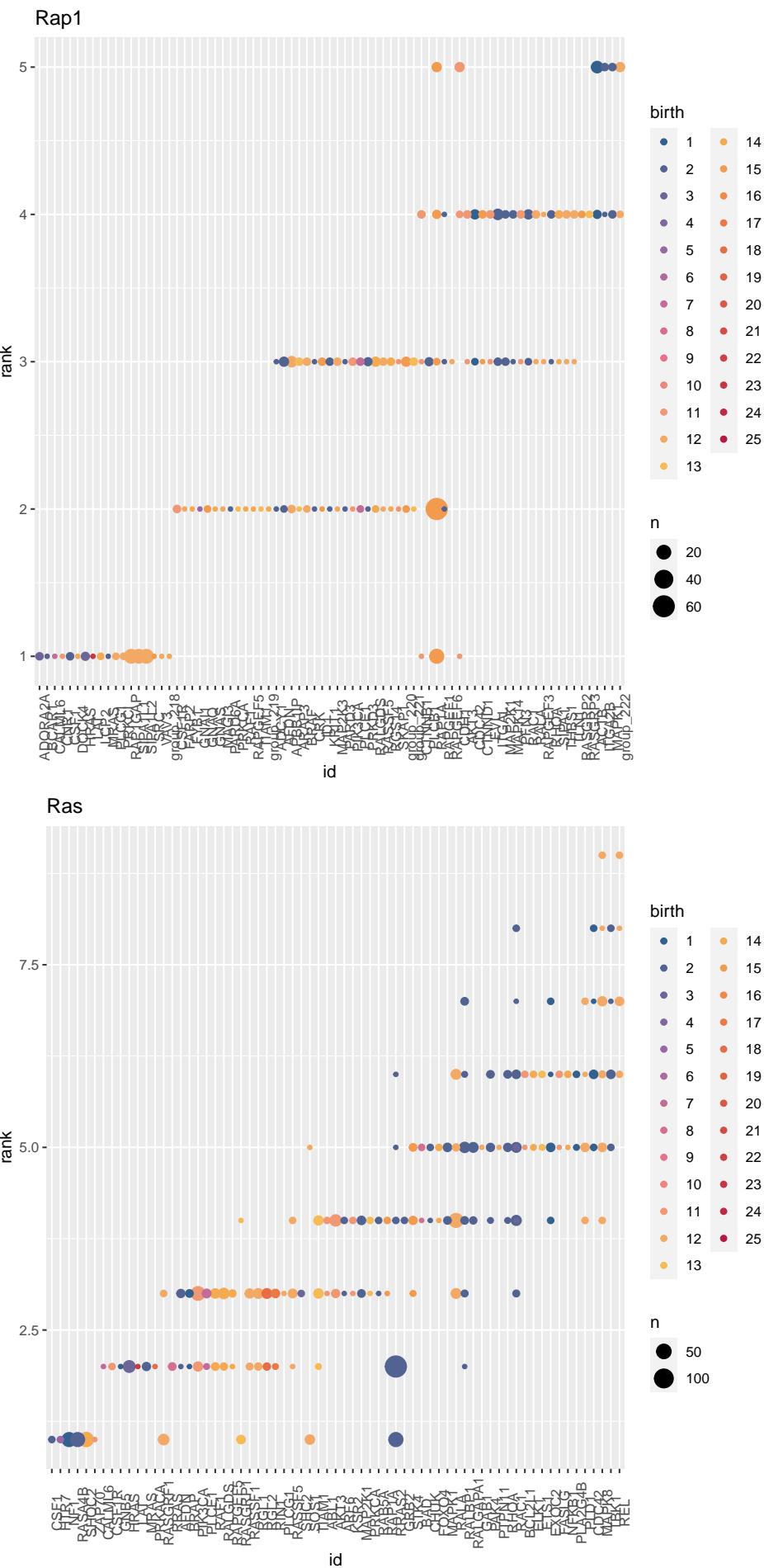


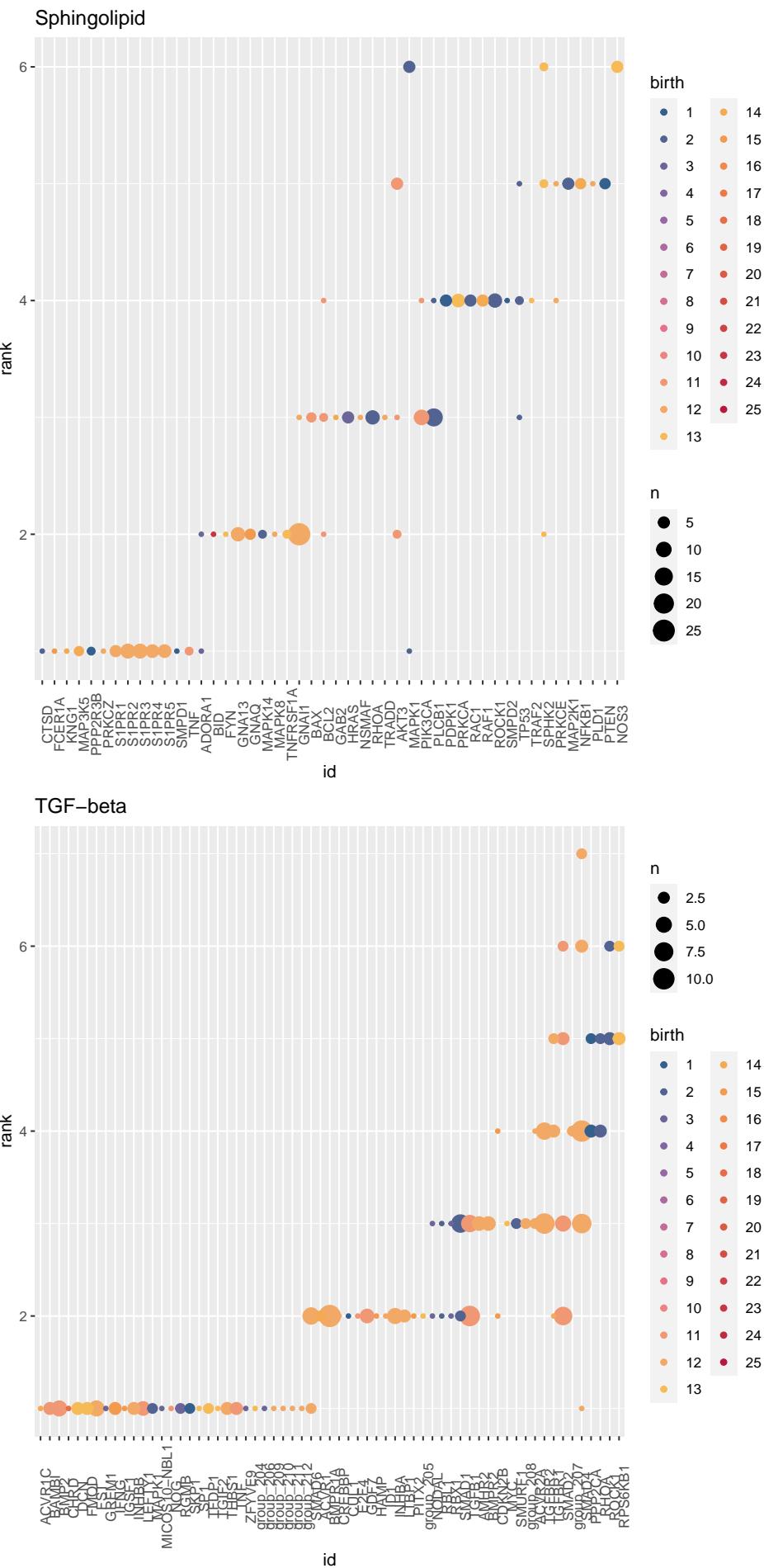
## Phospholipase D

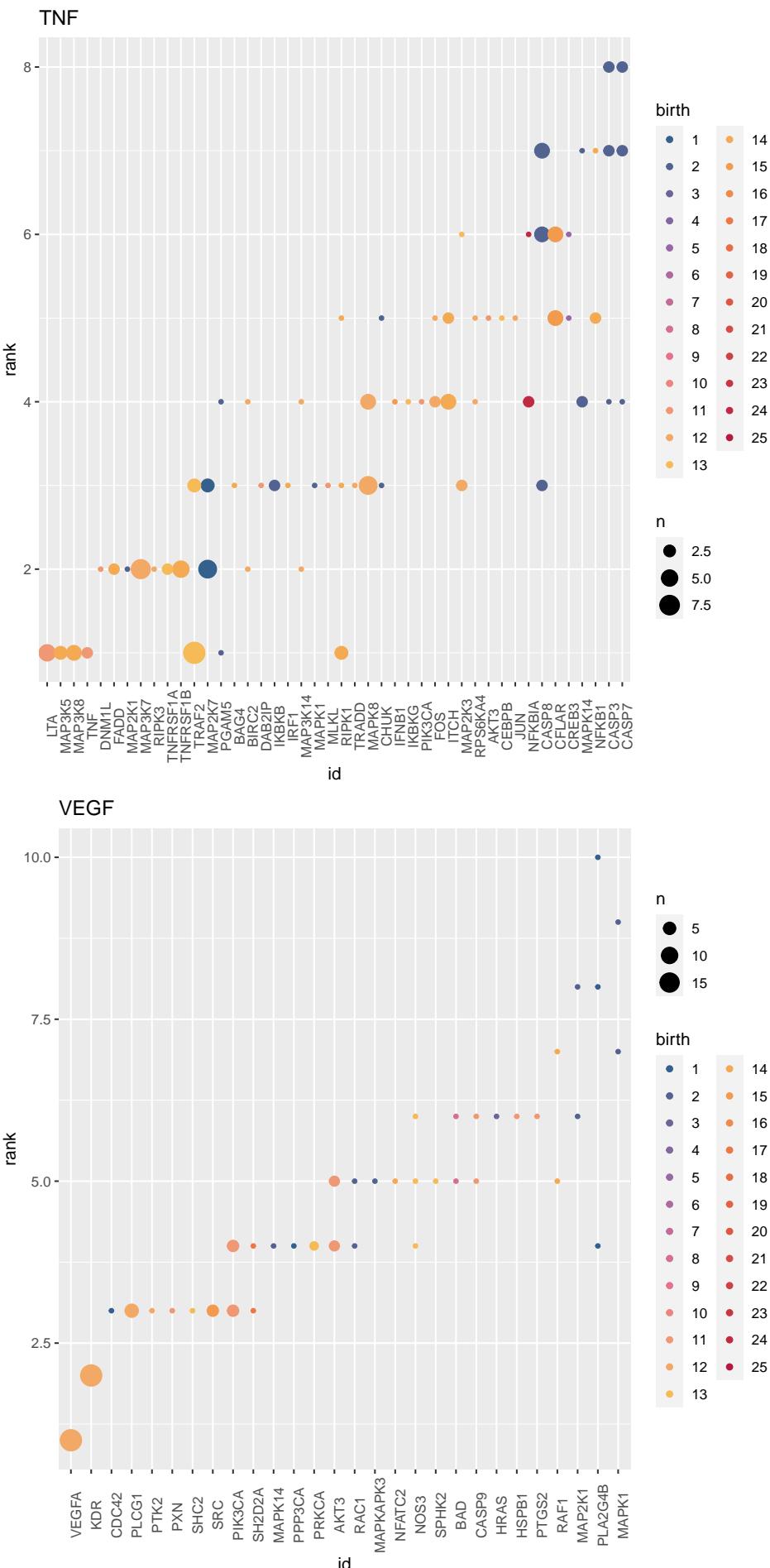


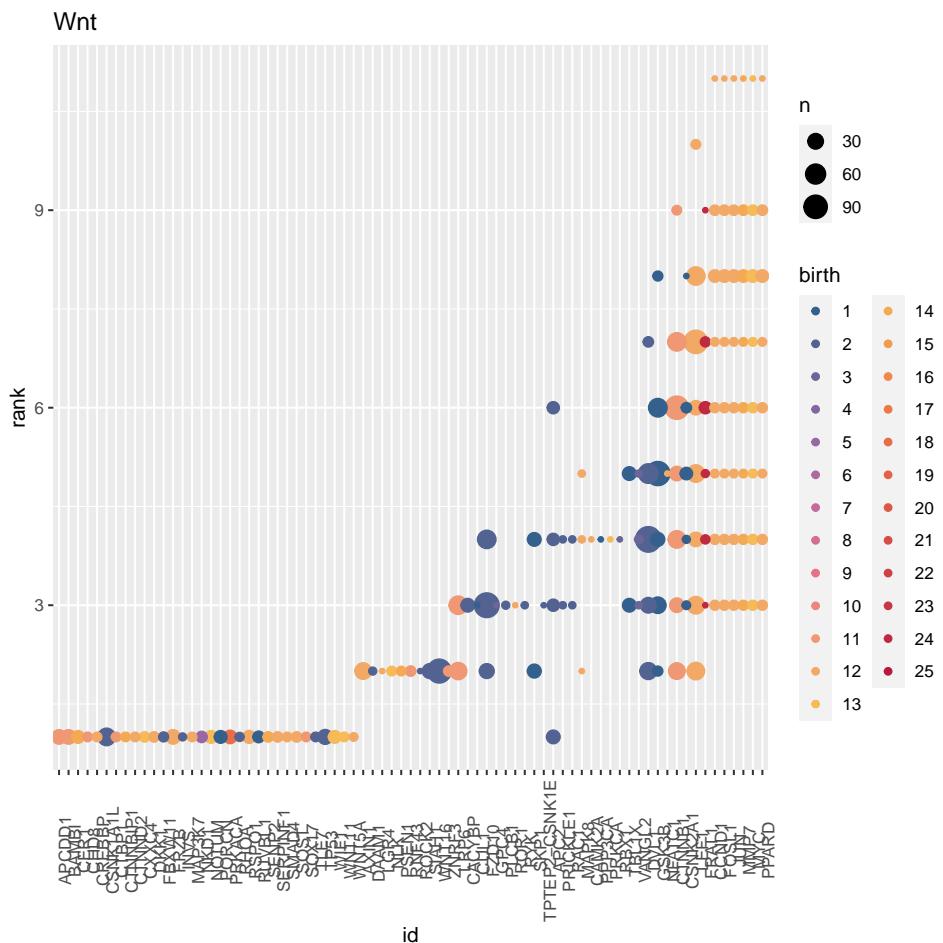
## PI3K-Akt









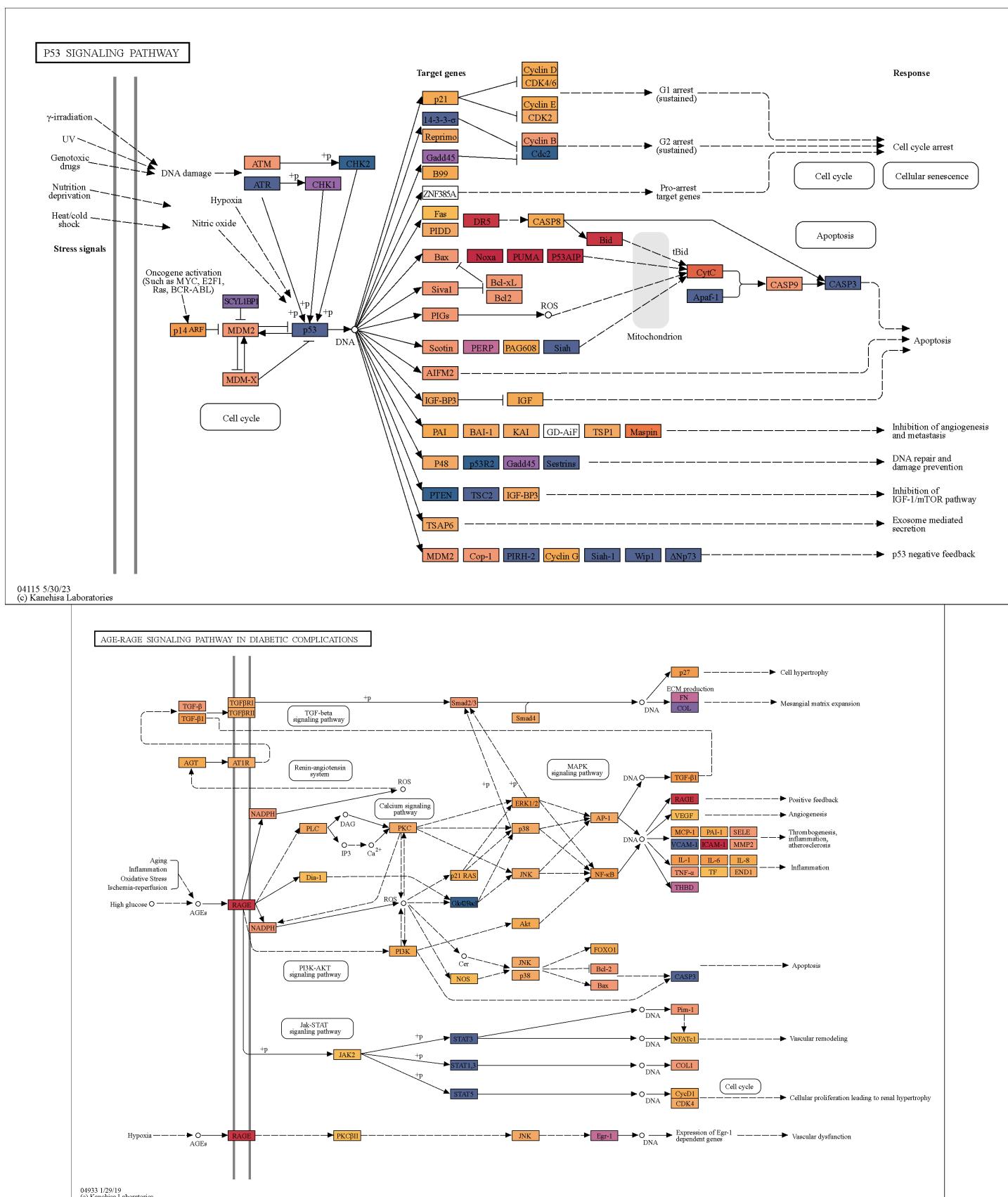


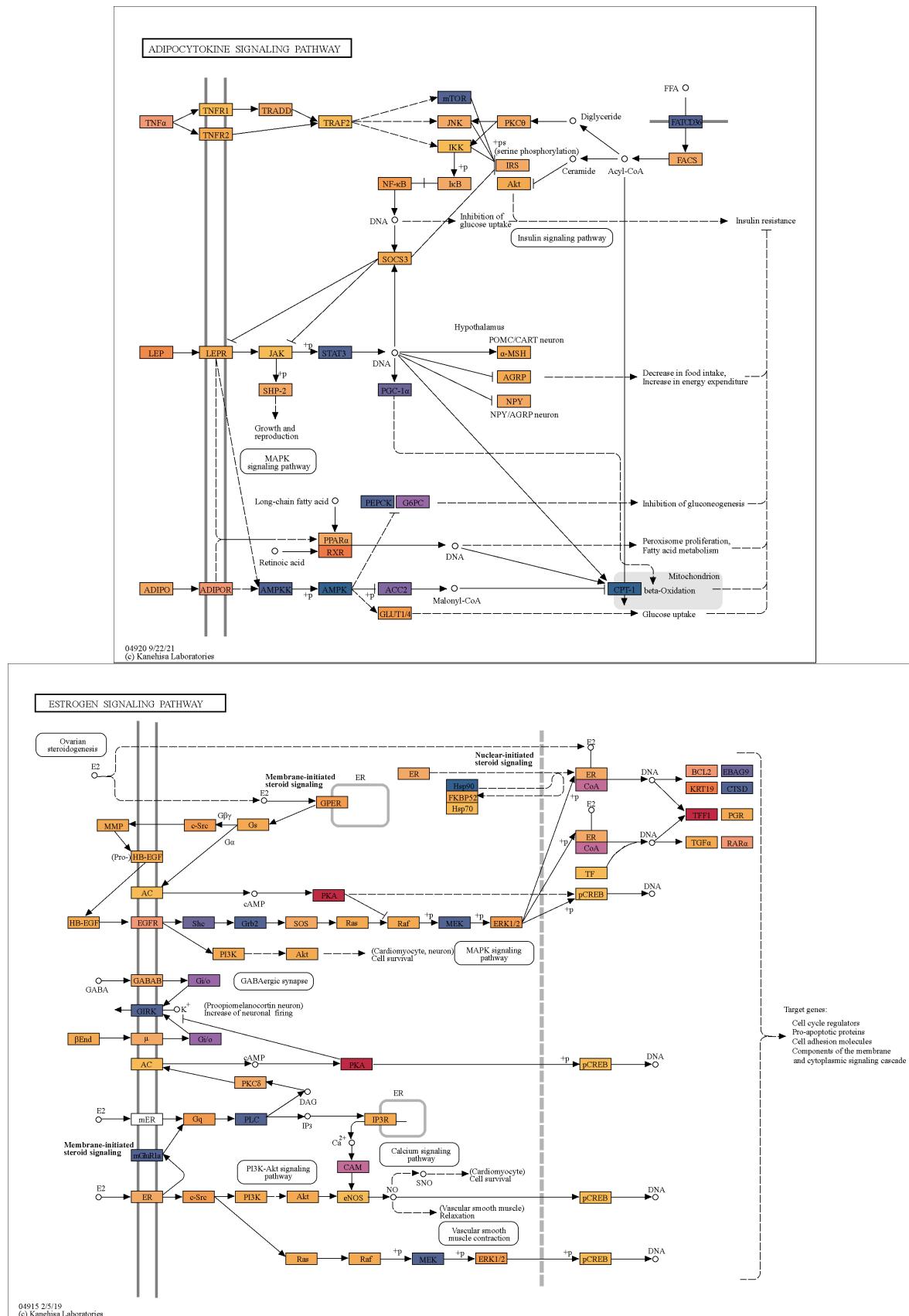
## 6.4 Article 1 - Suppl. Data 3

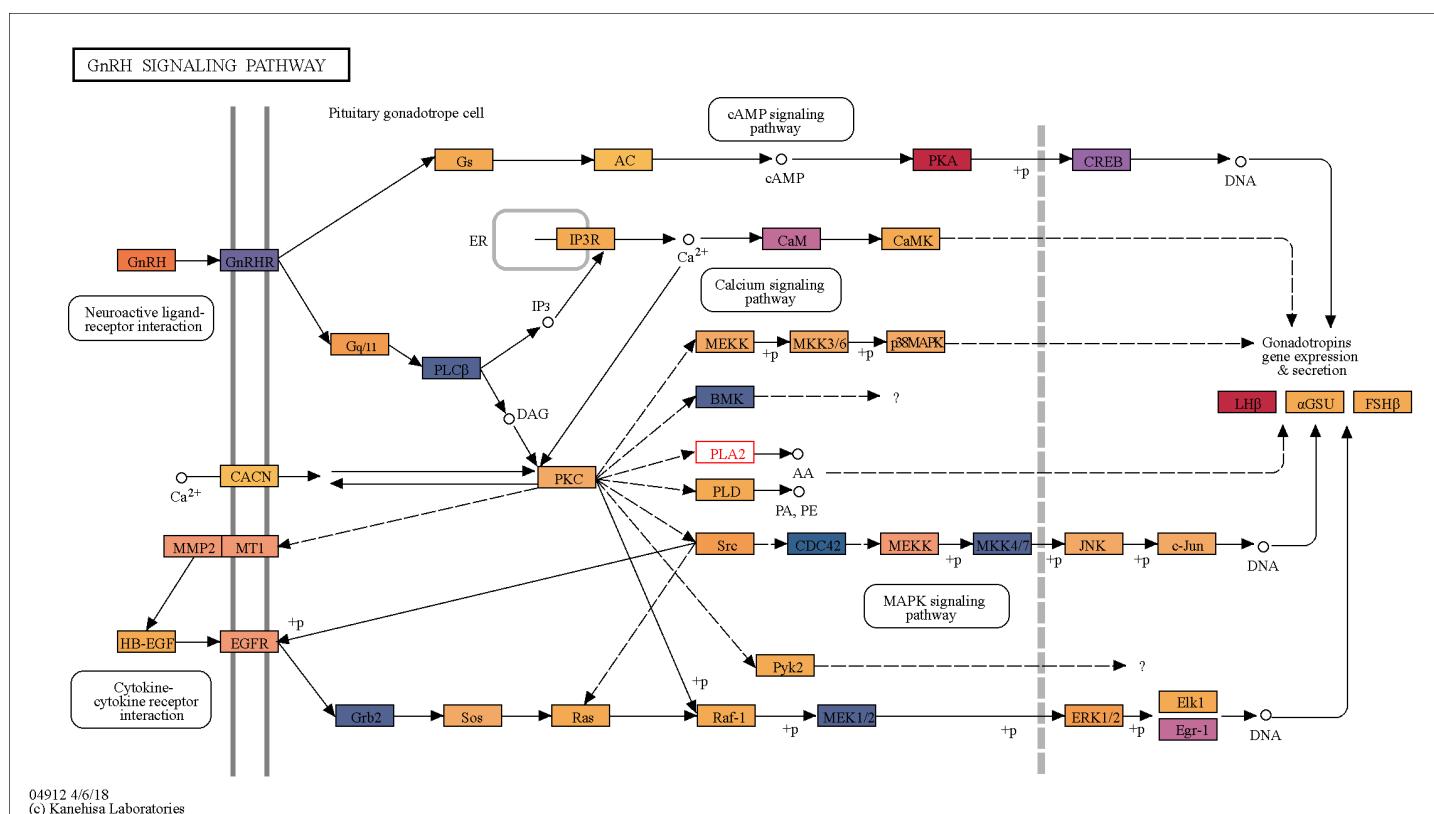
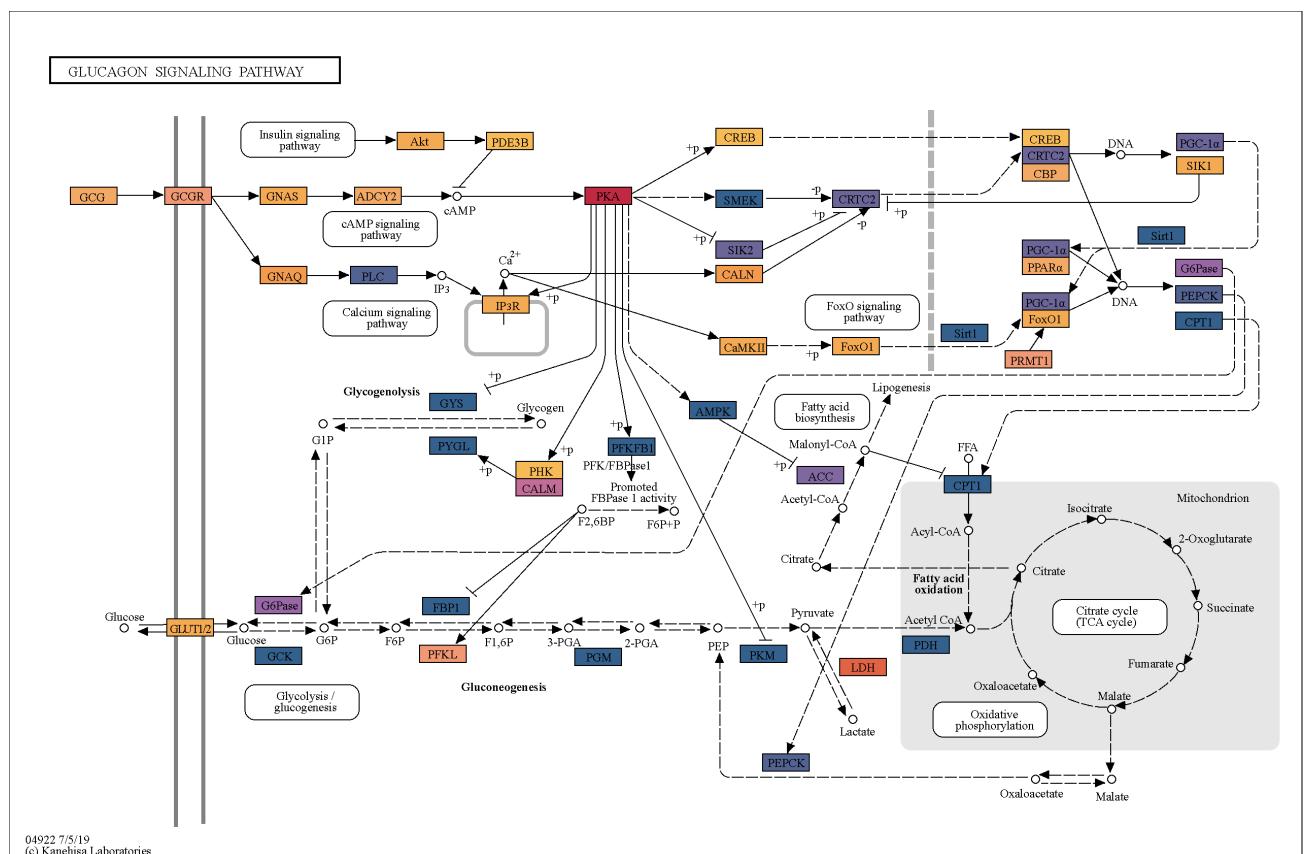
### Suppl. Data 3 - Colored KEGG pathways depending on the node of birth of each protein

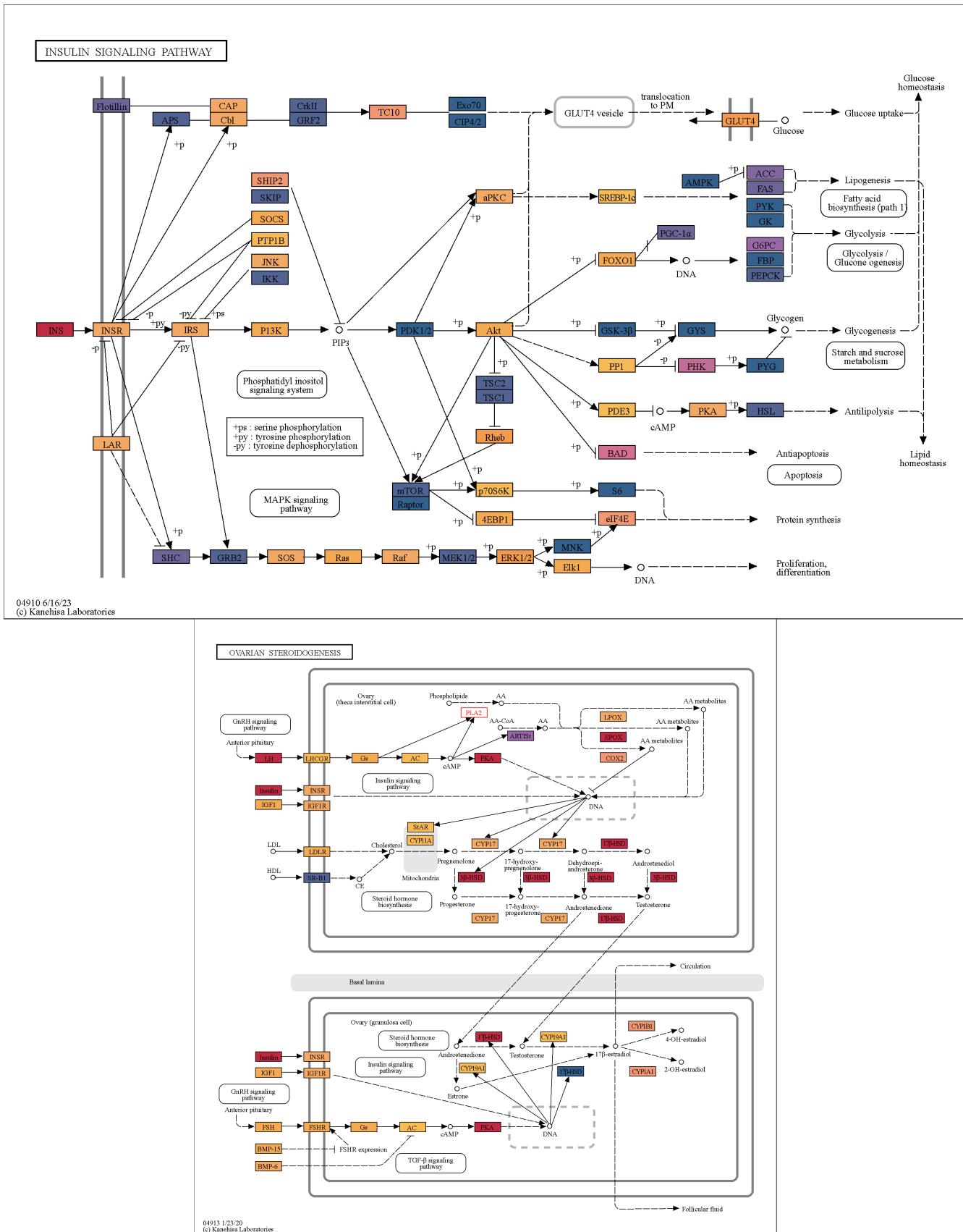
Legend : Each color represents a clade. The white rectangles correspond to the genes for which we have not been able to determine the node of birth due to lack of information about the gene. The KEGG legend is available here : [https://www.genome.jp/kegg/document/help\\_pathway.html](https://www.genome.jp/kegg/document/help_pathway.html).

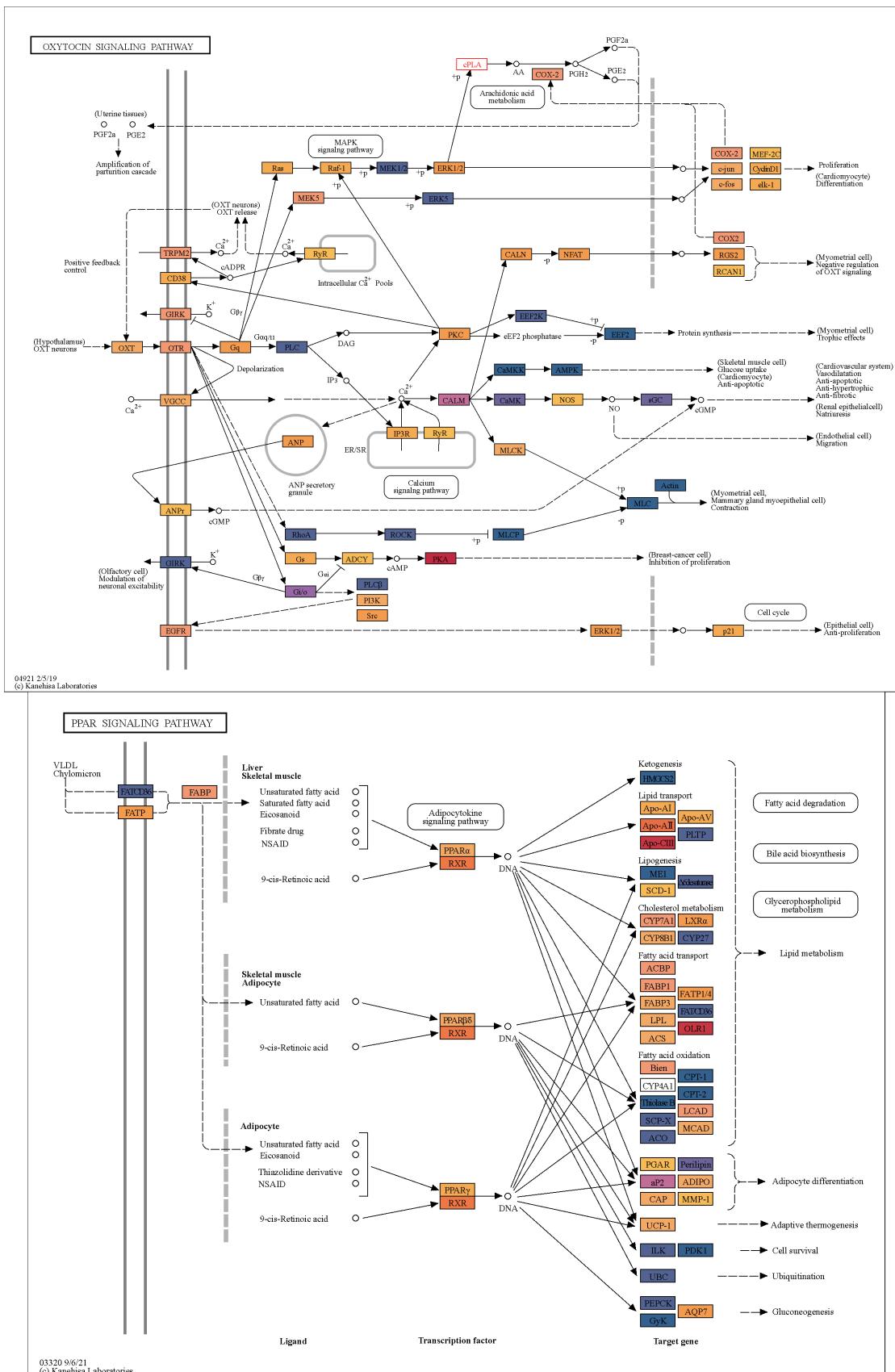
The pathways are in the following order : p53, AGE-RAGE, Adipocytokine, Estrogen, Glucagon, GnRH, Insulin, Ovarian steroidogenesis, Oxytocin, PPAR, Prolactin, Relaxin, Thyroid hormone, B cell receptor, C-type lectin receptor, Chemokine, FC epsilon RI, IL-17, NOD-like receptor, RIG-I-like receptor, T cell receptor, Toll-like receptor, Neurotrophin, AMPK, Apelin, Calcium, cAMP, cGMP-PKG, ErbB, FoxO, Hedgehog, HIF-1, Hippo, JAK-STAT, MAPK, mTOR, NF-Kappa B, Notch, Phospholipase D, PI3K-Akt, Rap1, Ras, Sphingolipid, TGF-Beta, TNF , VEGF, Wnt.

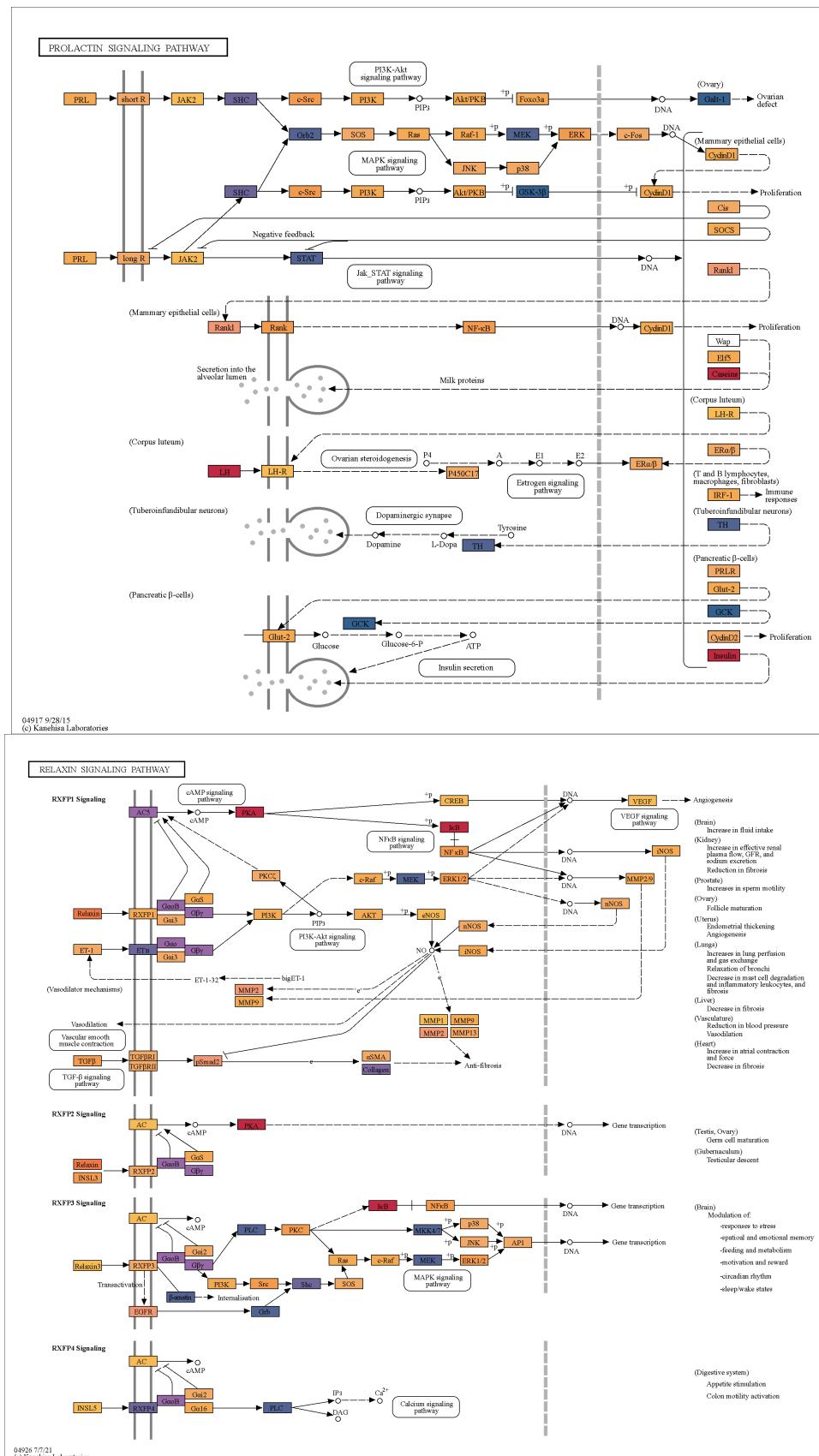


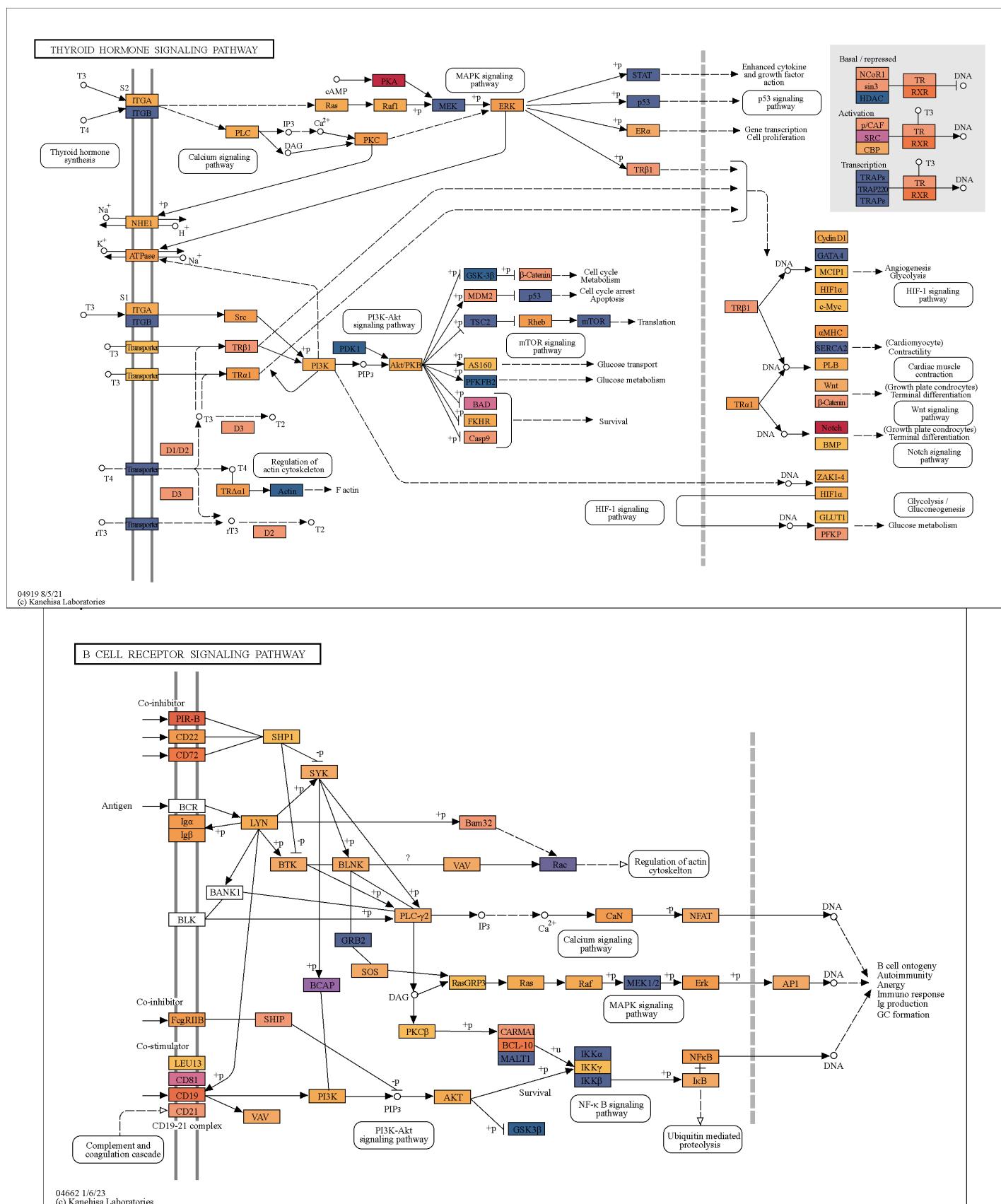


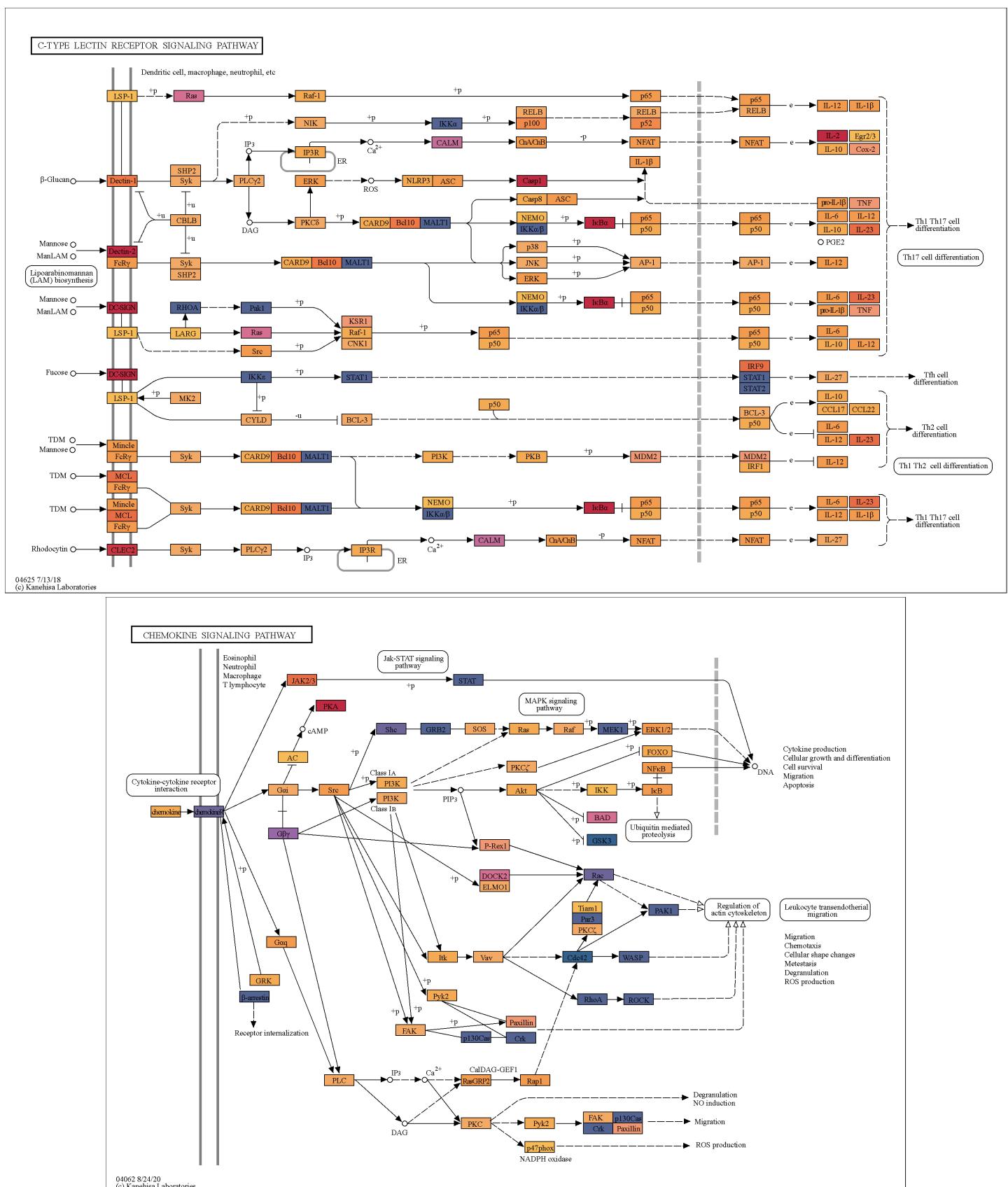


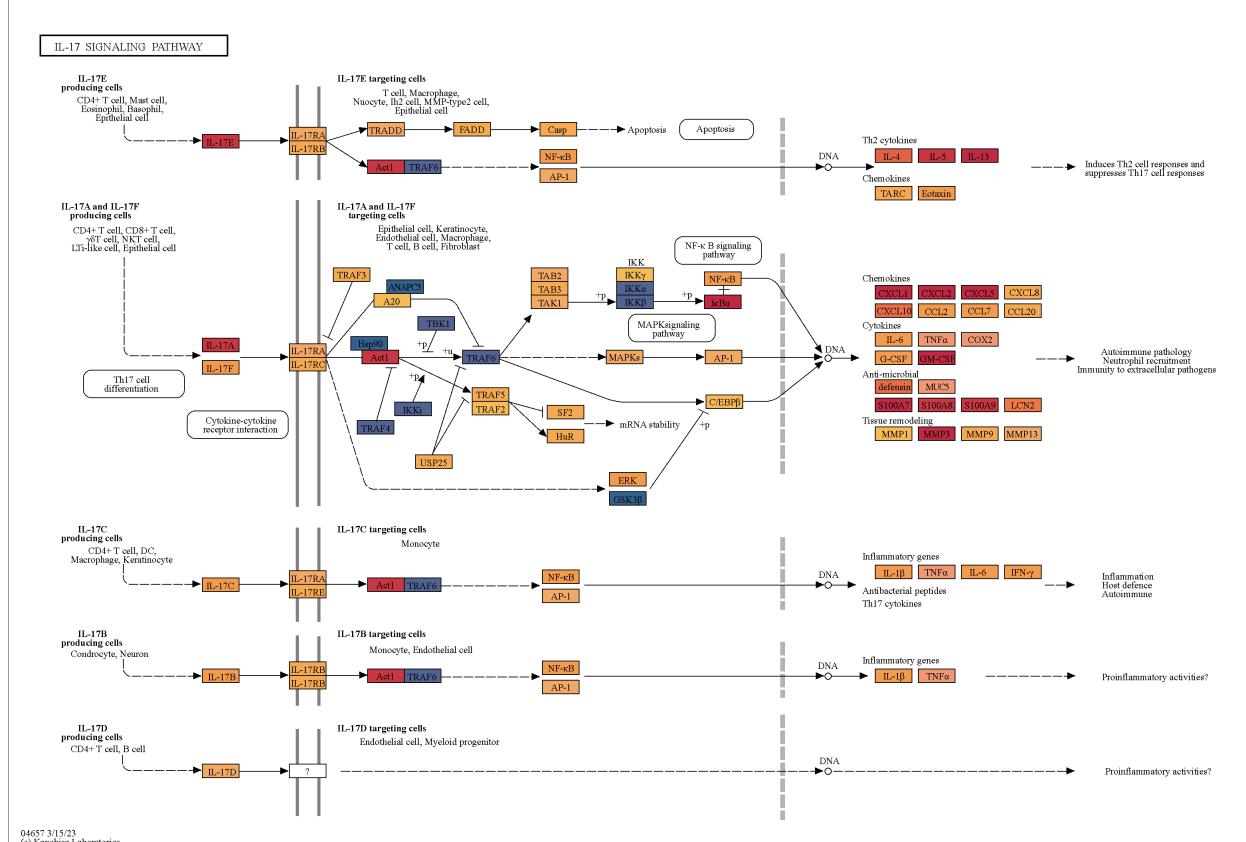
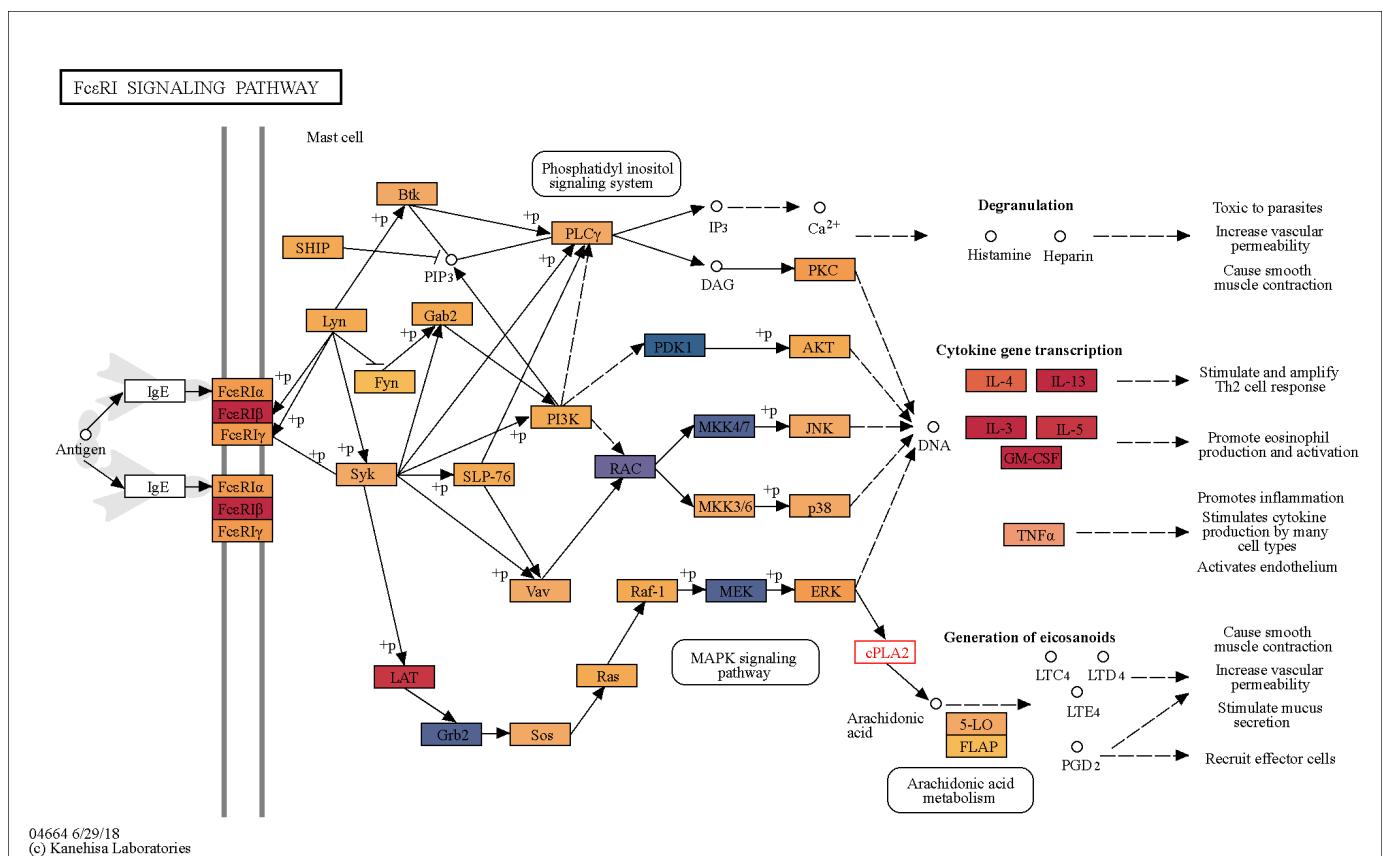


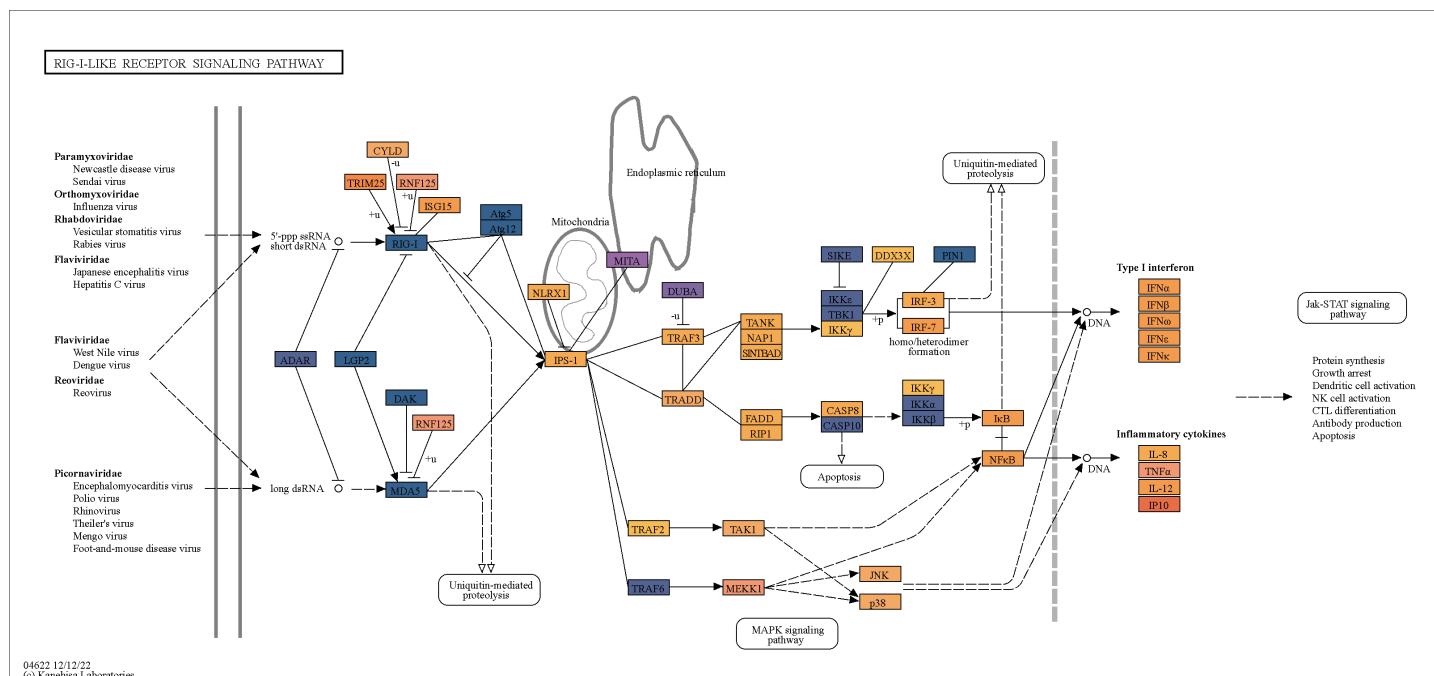
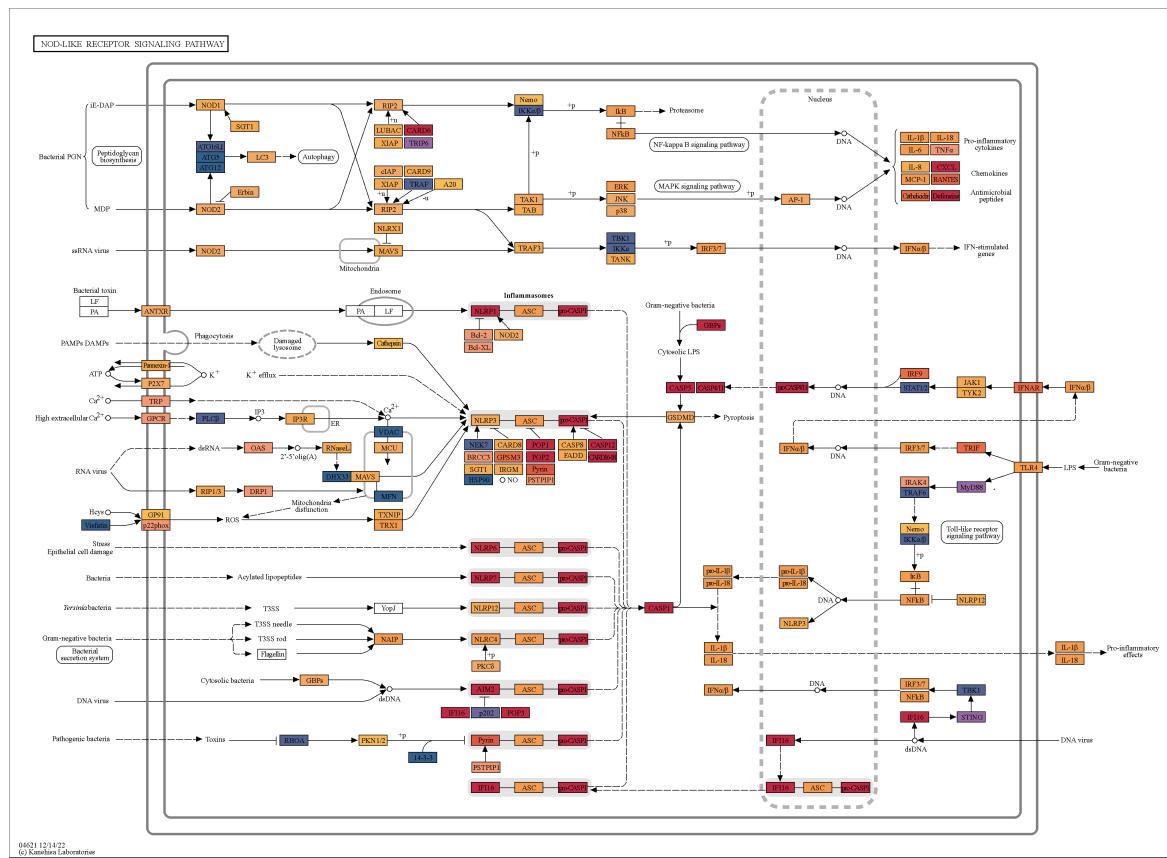


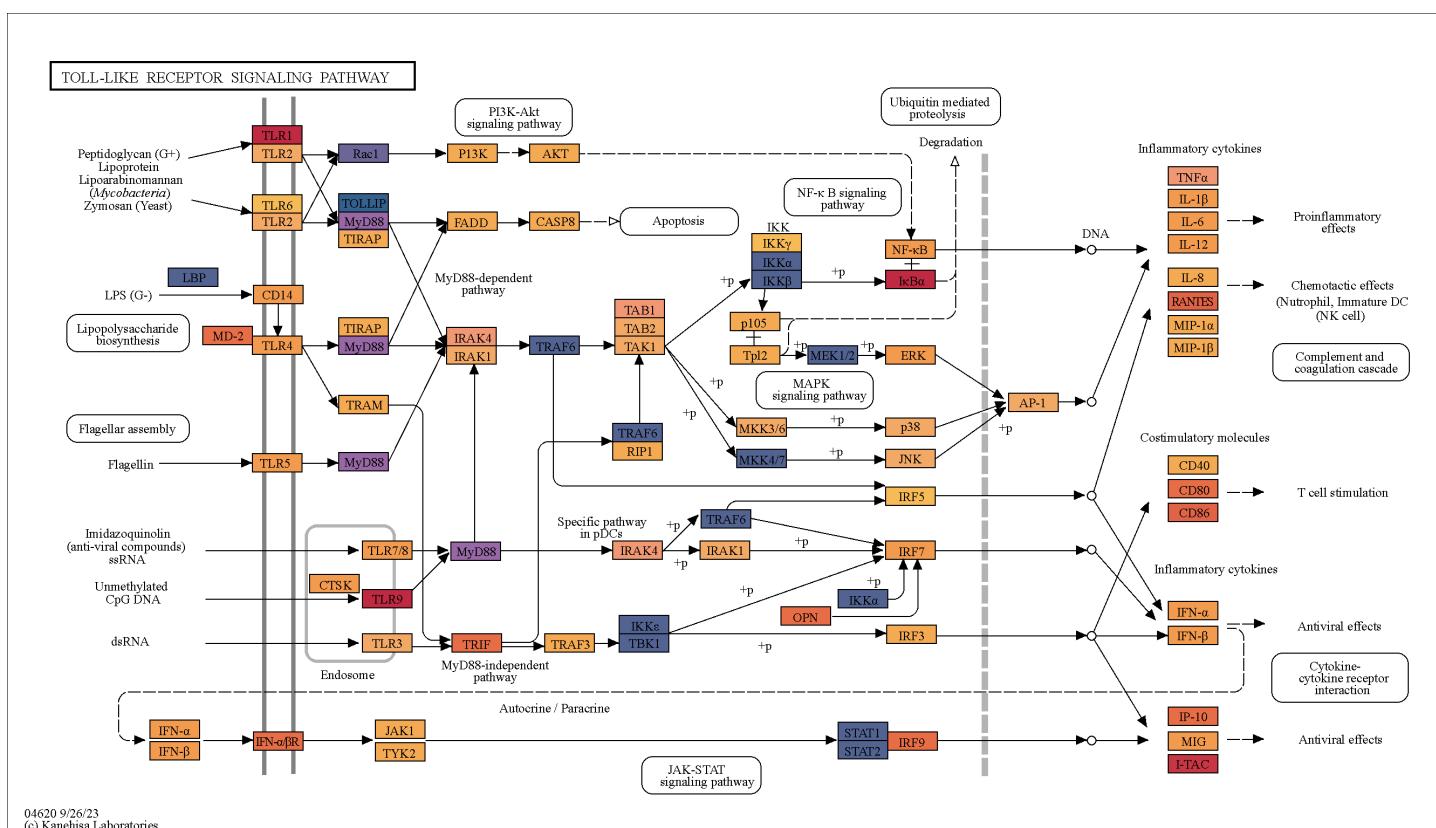
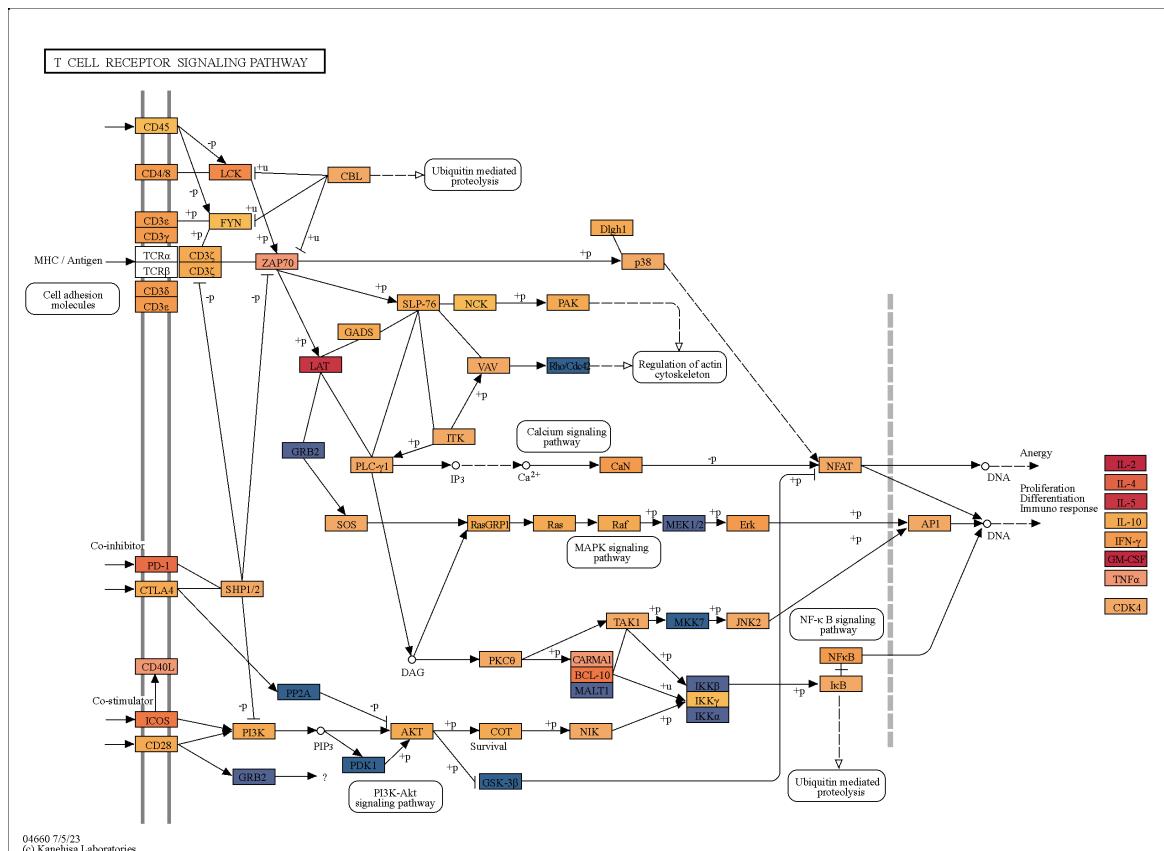


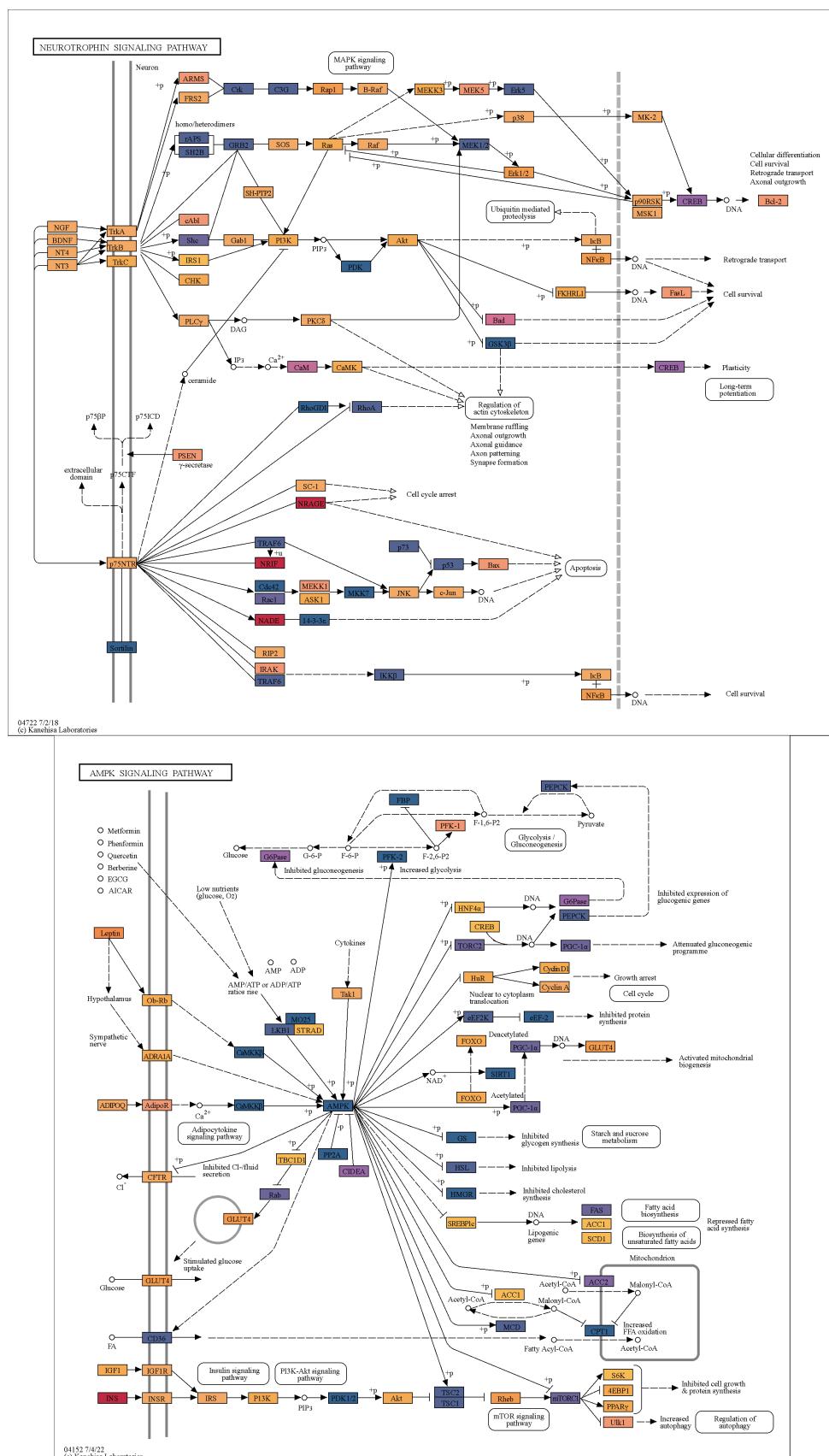


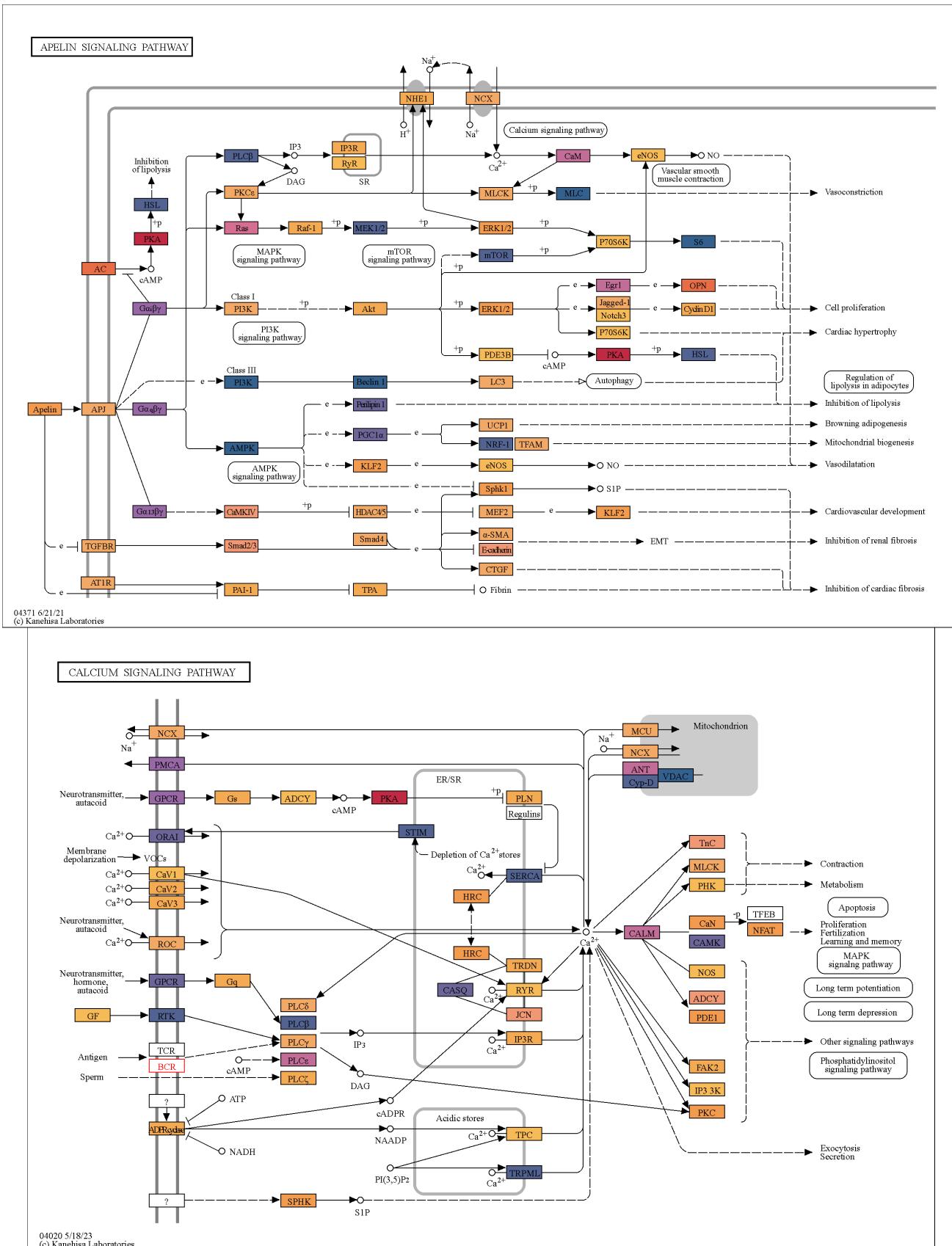


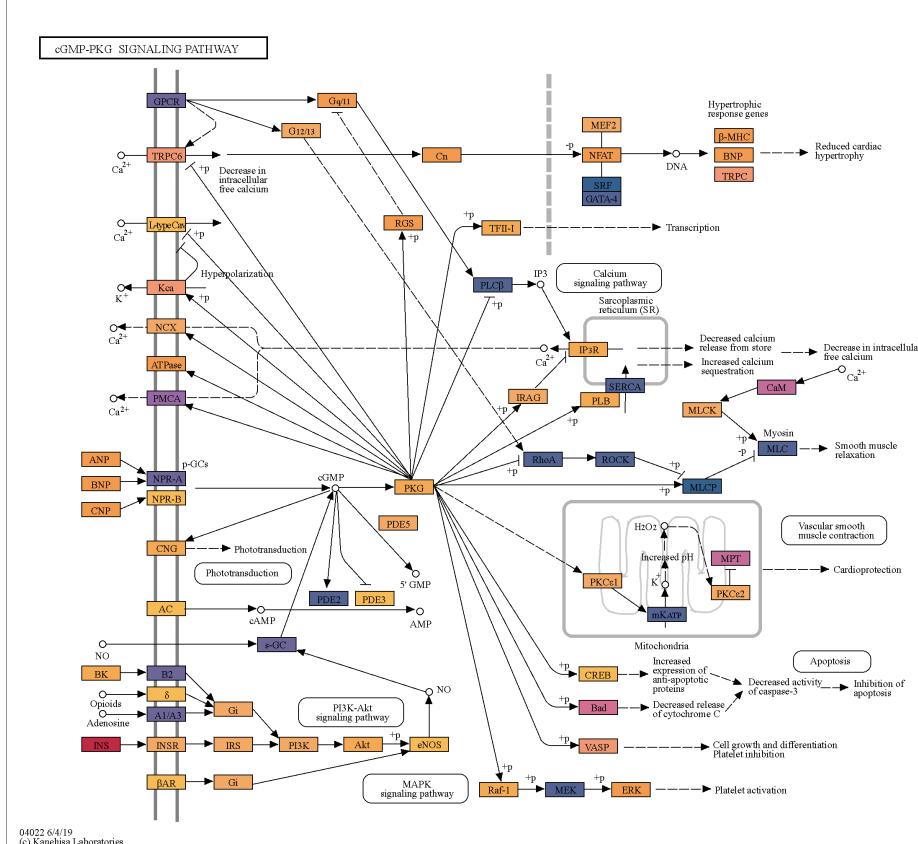
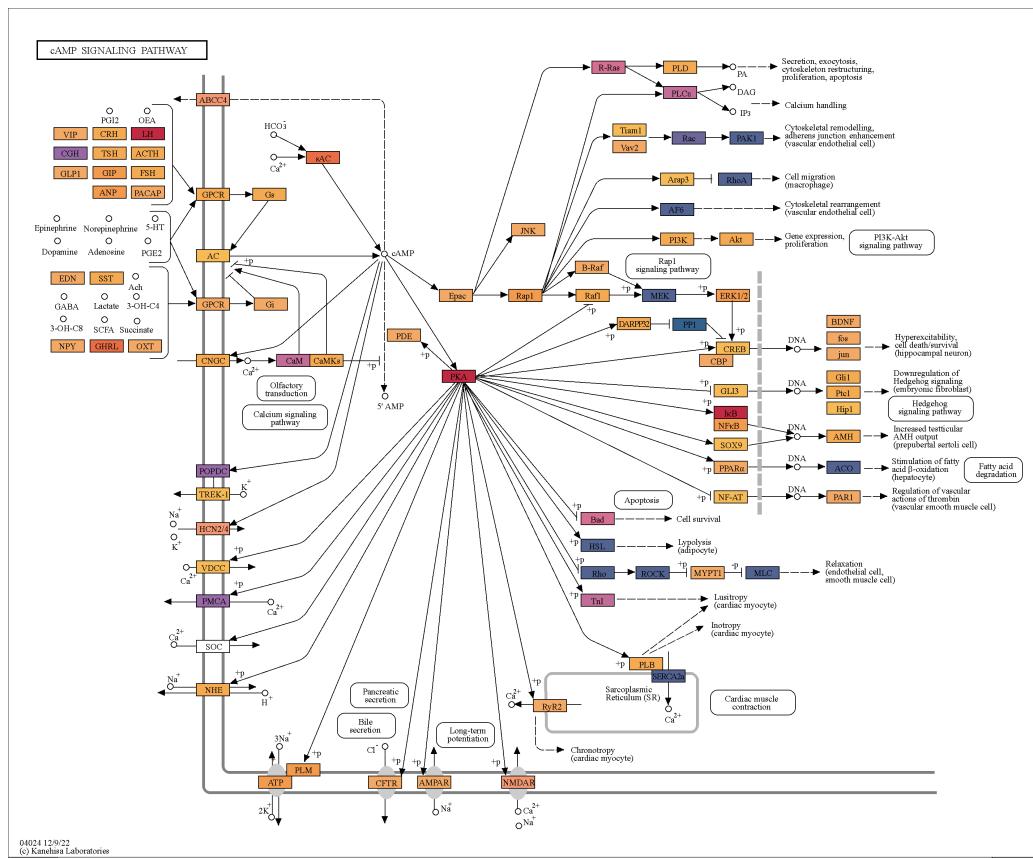


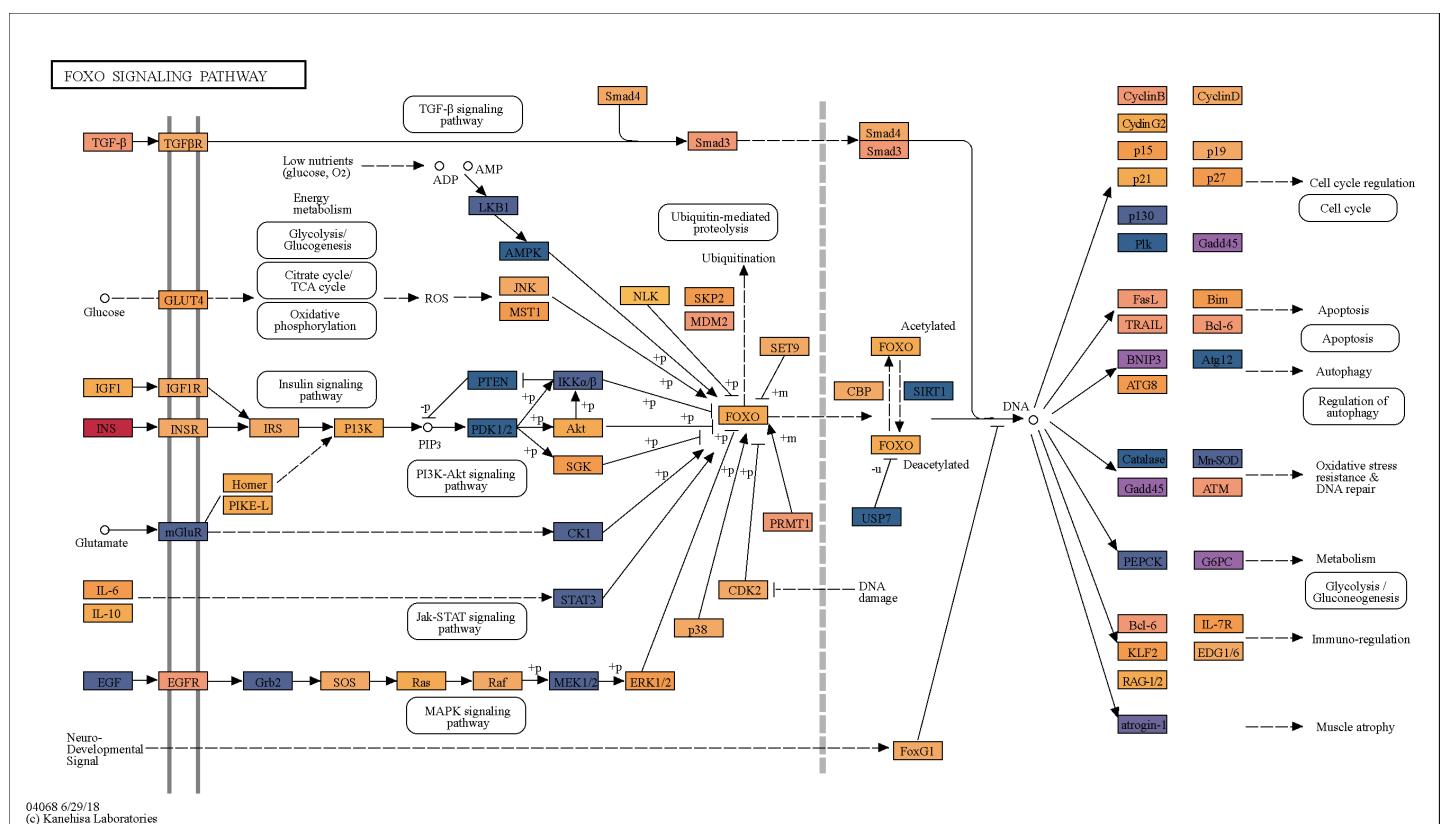
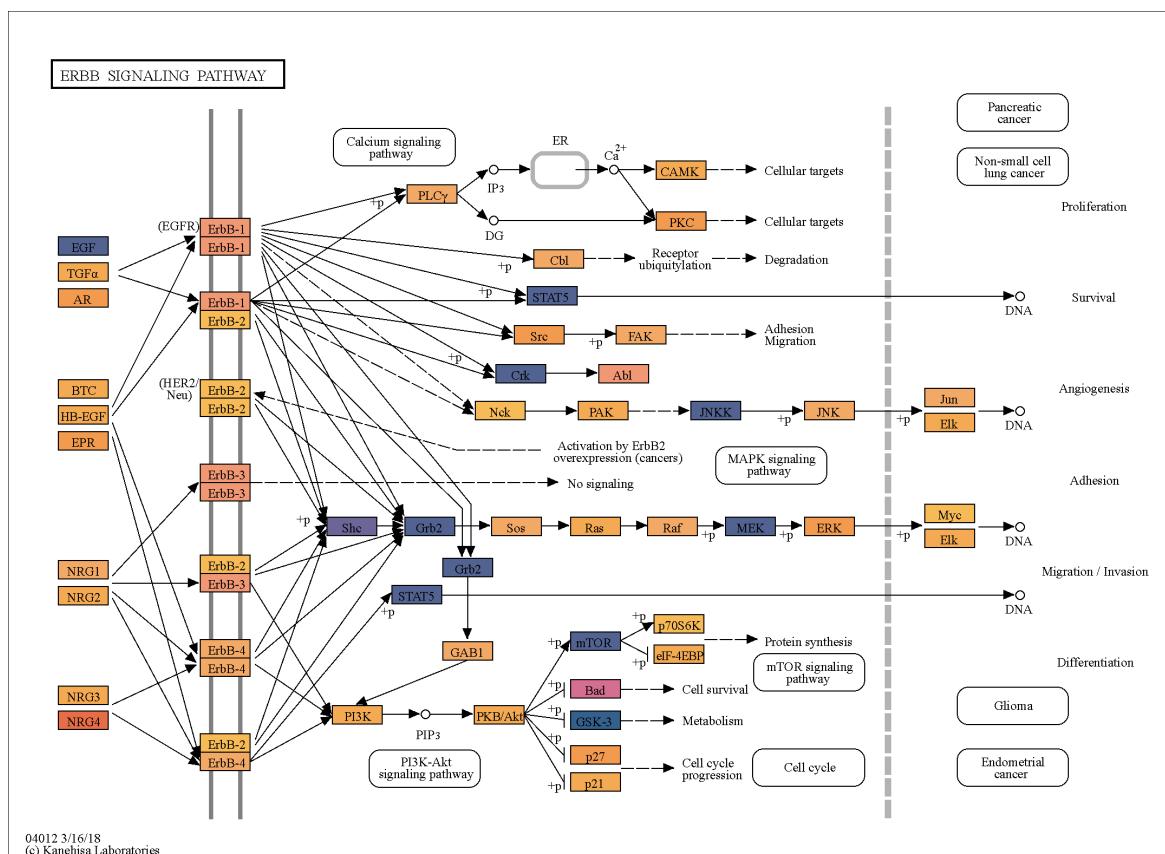


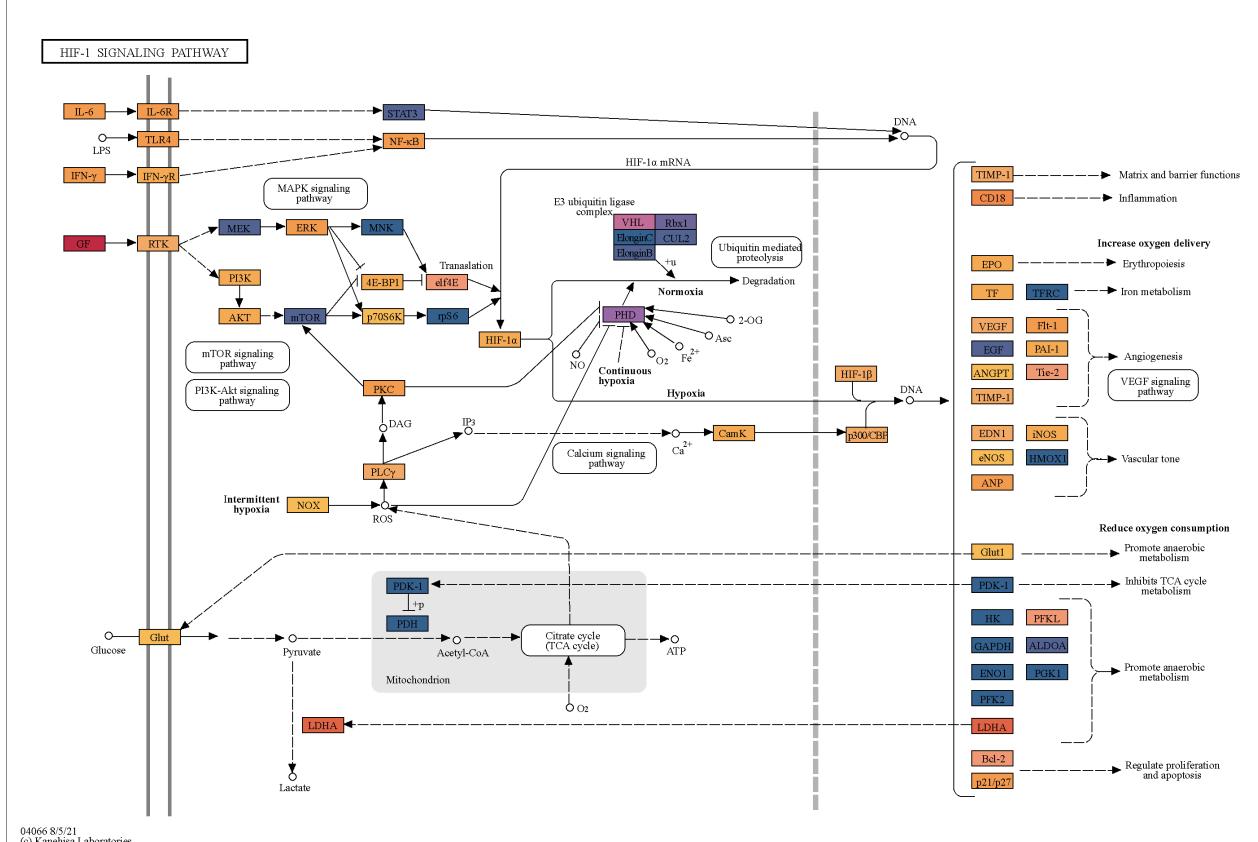
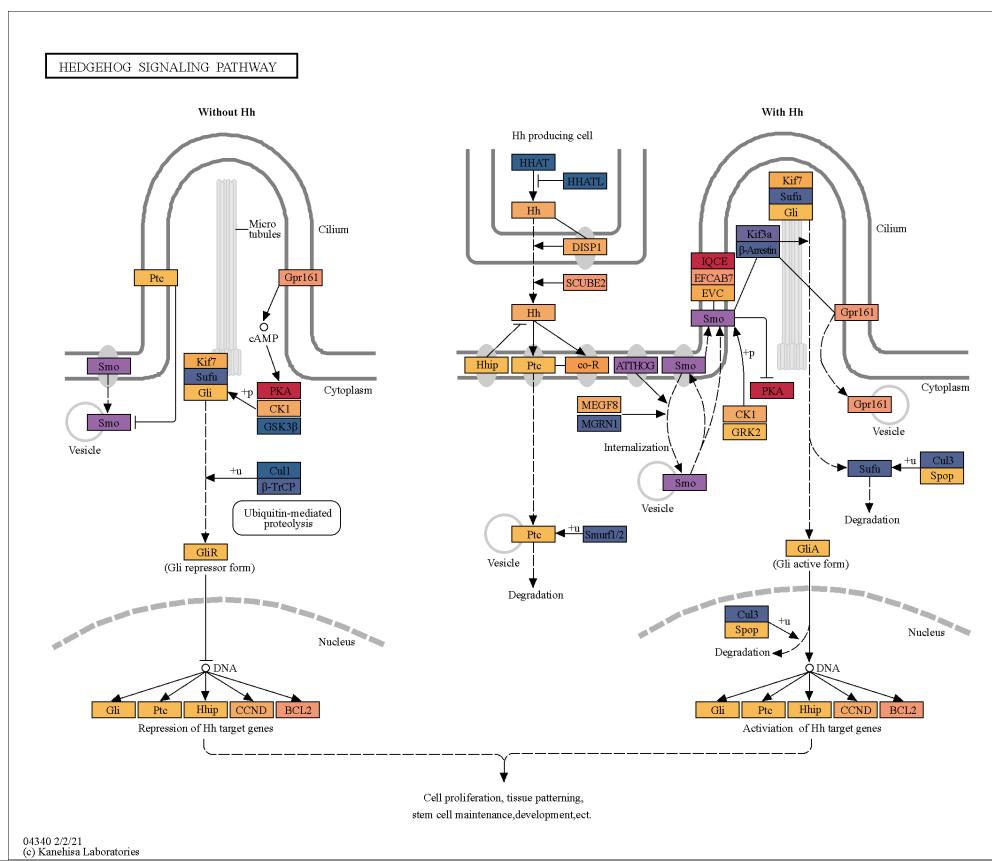


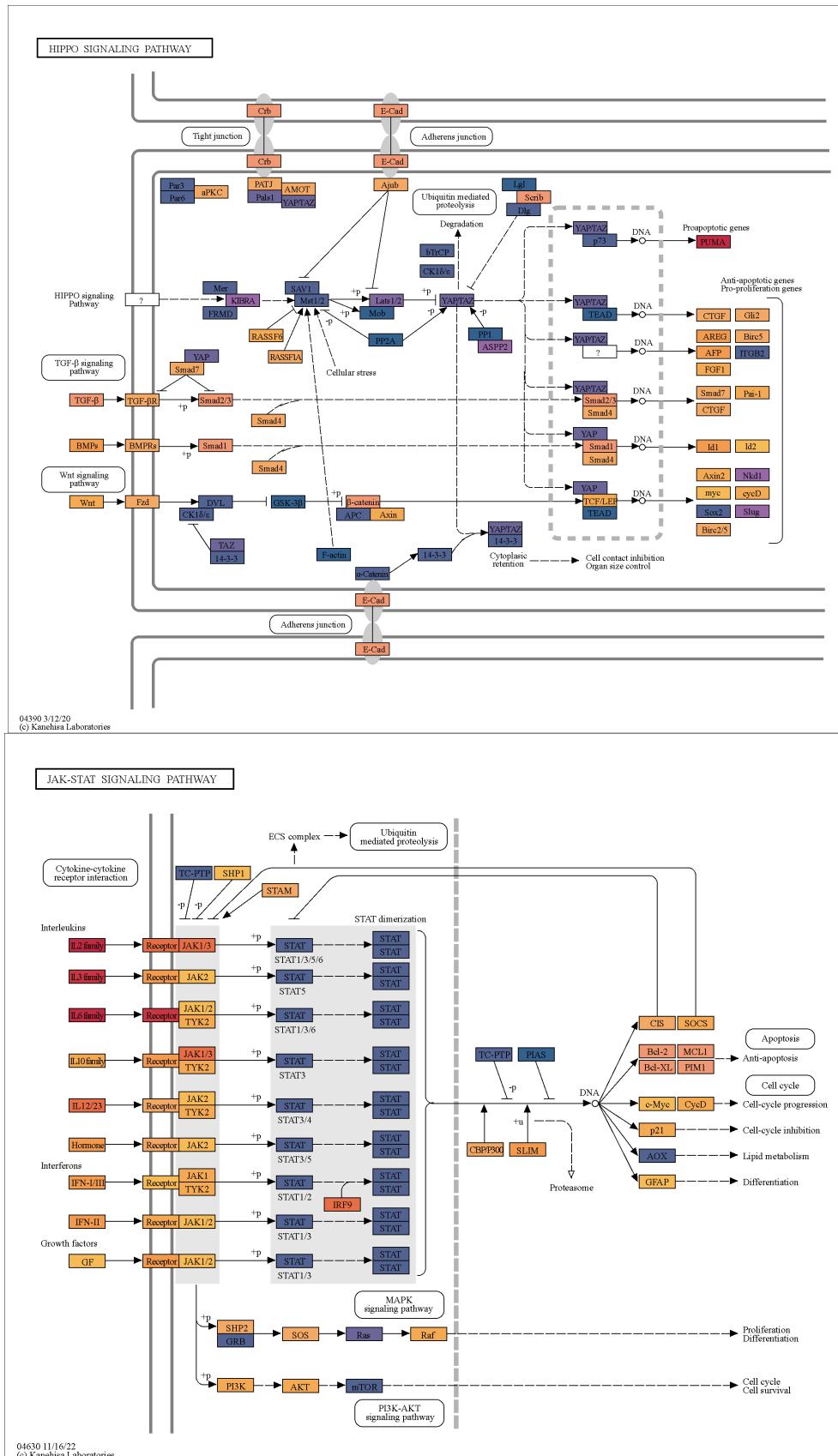


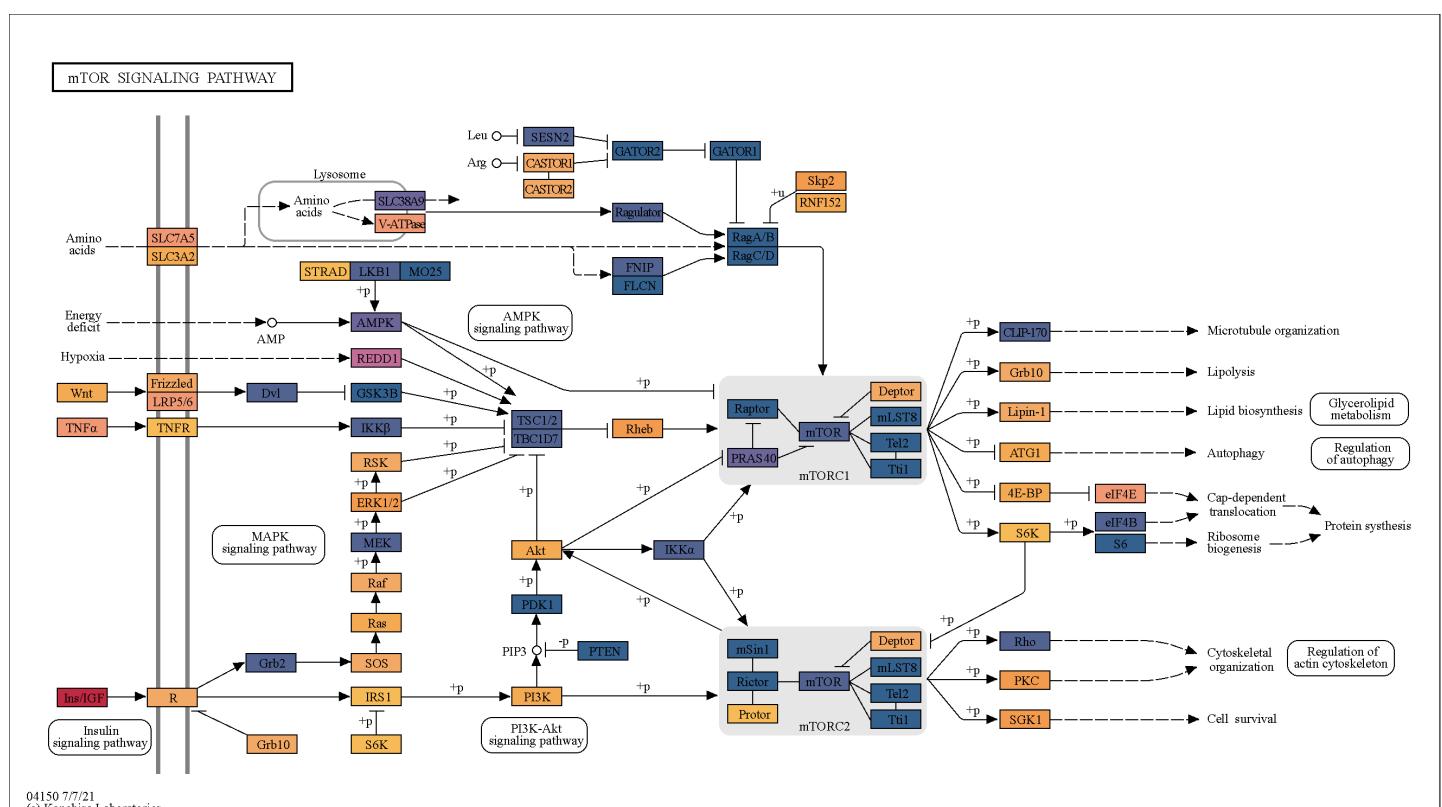
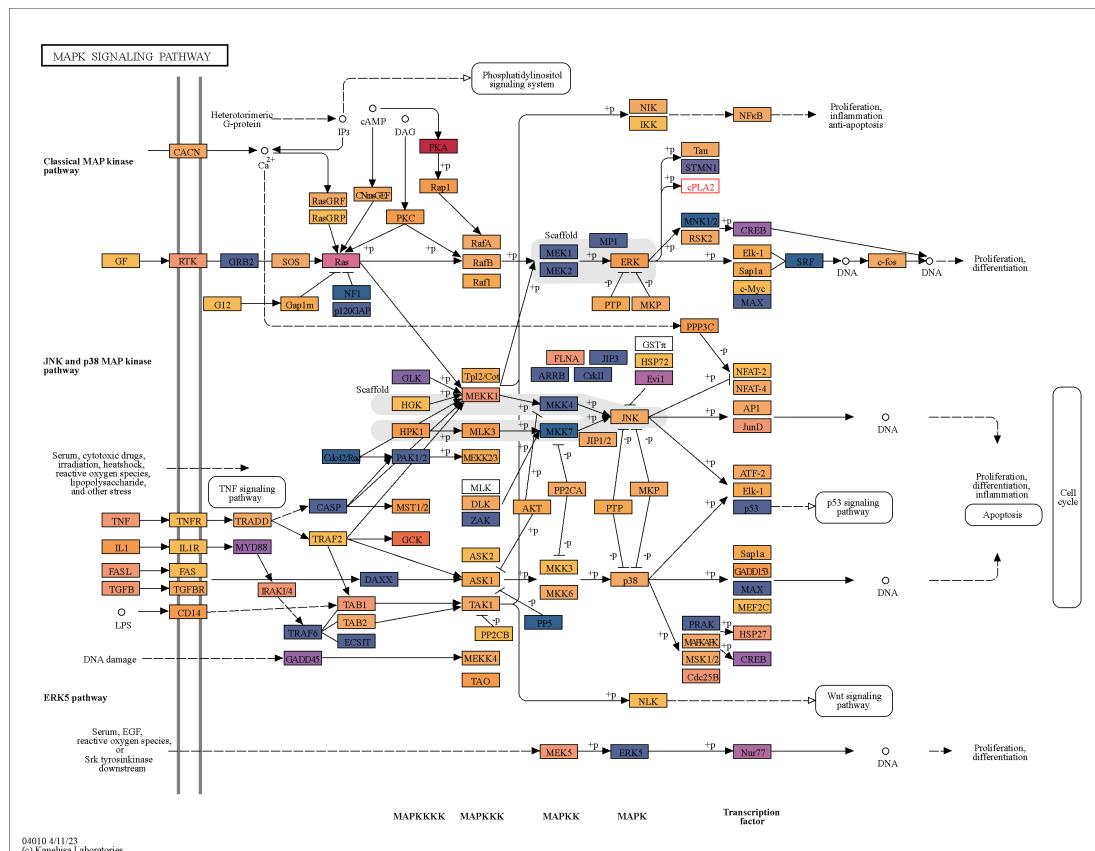


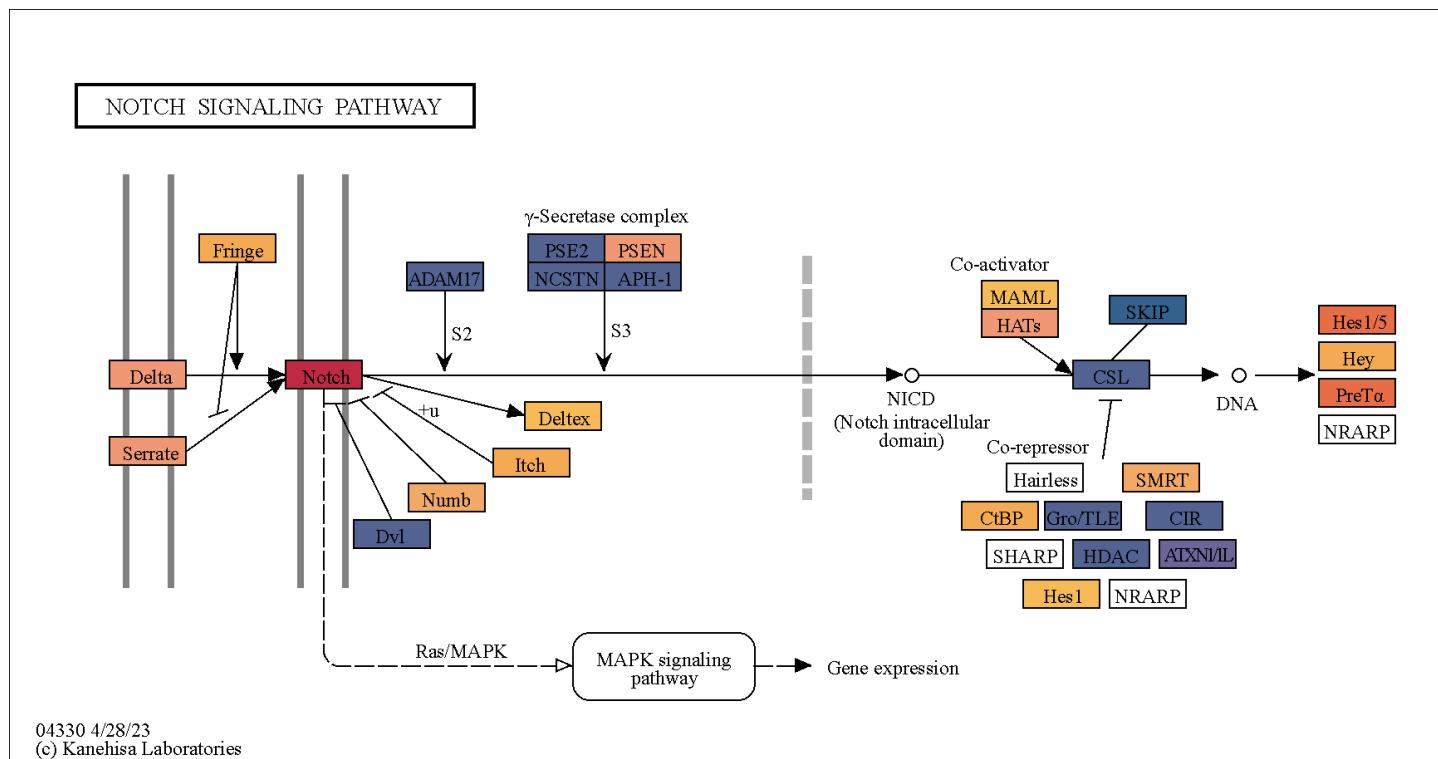
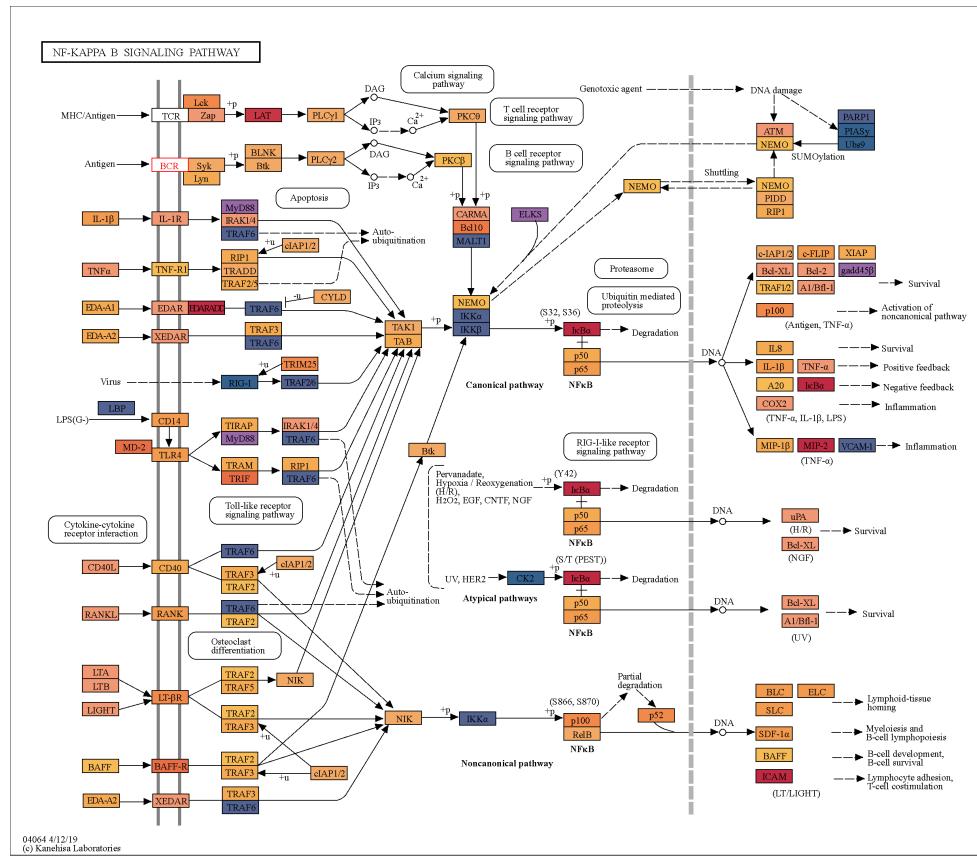


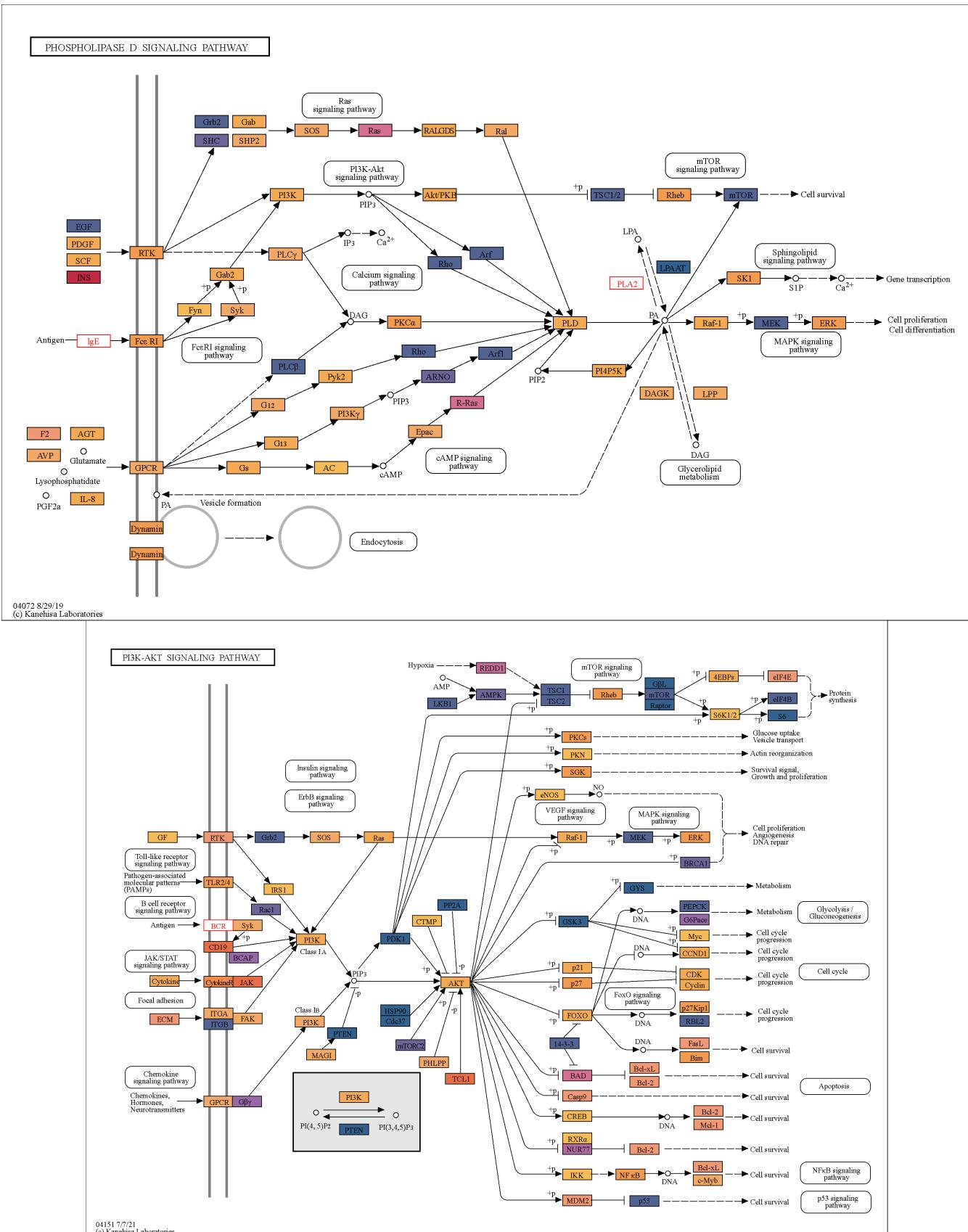


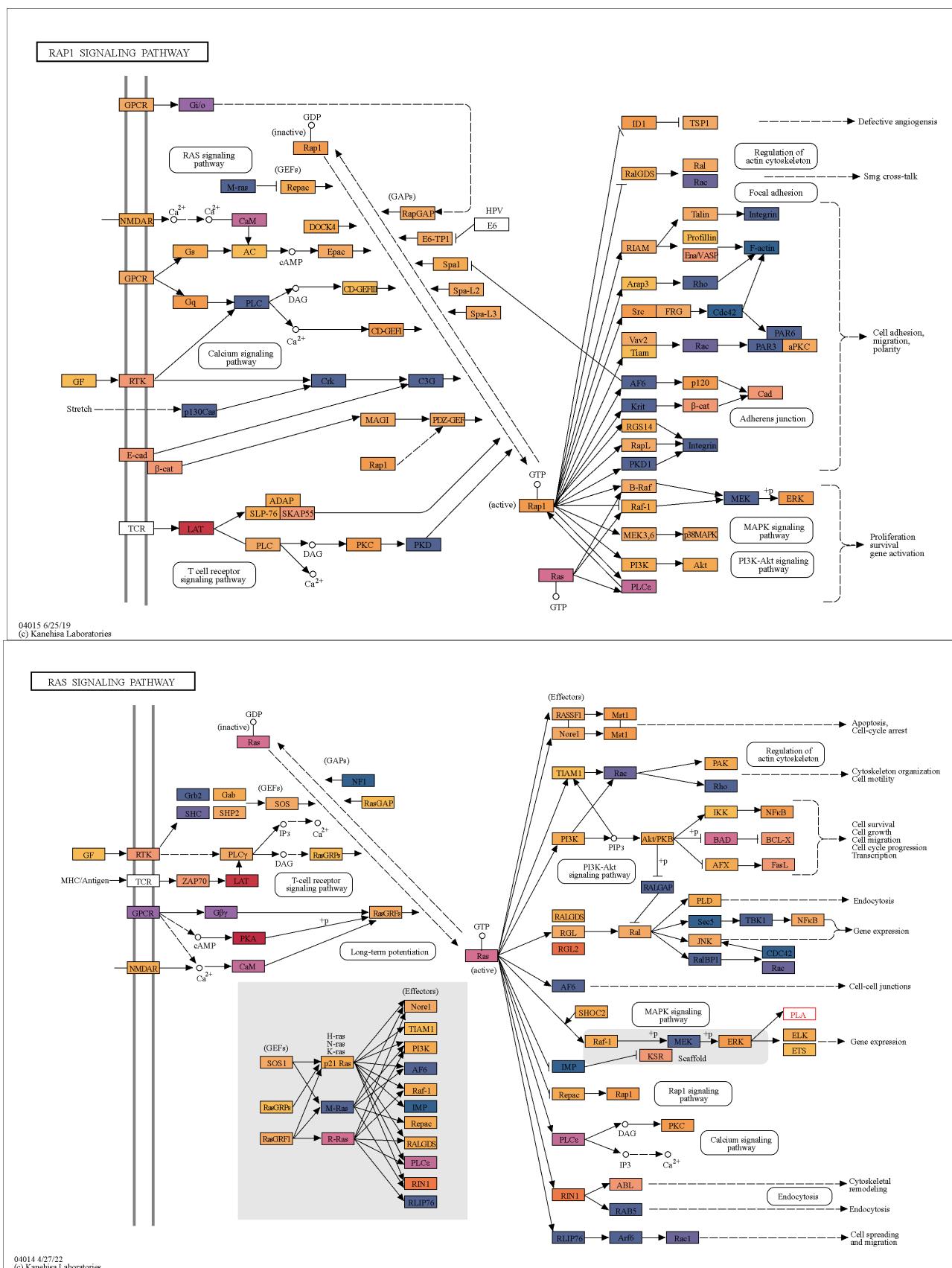


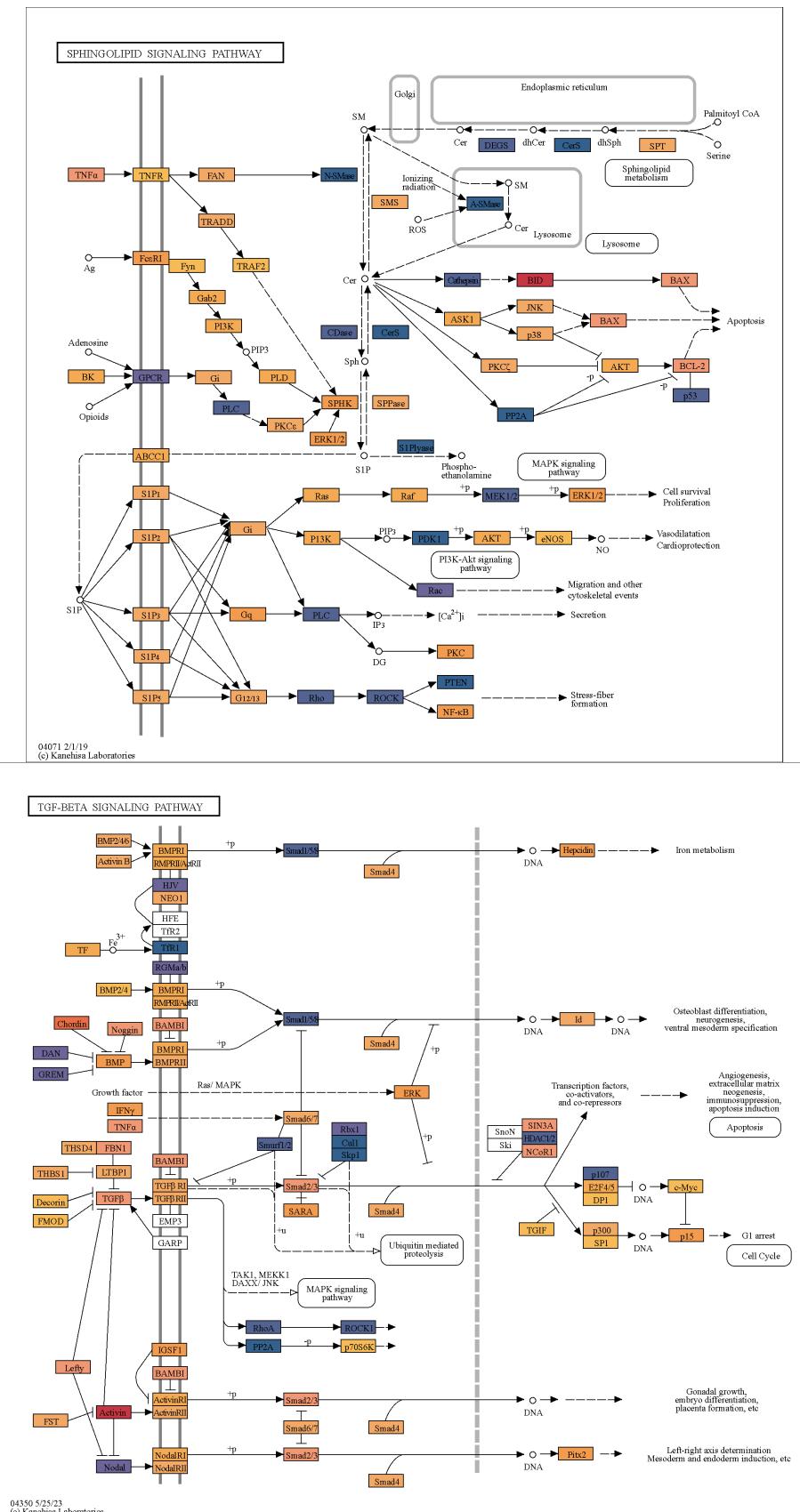


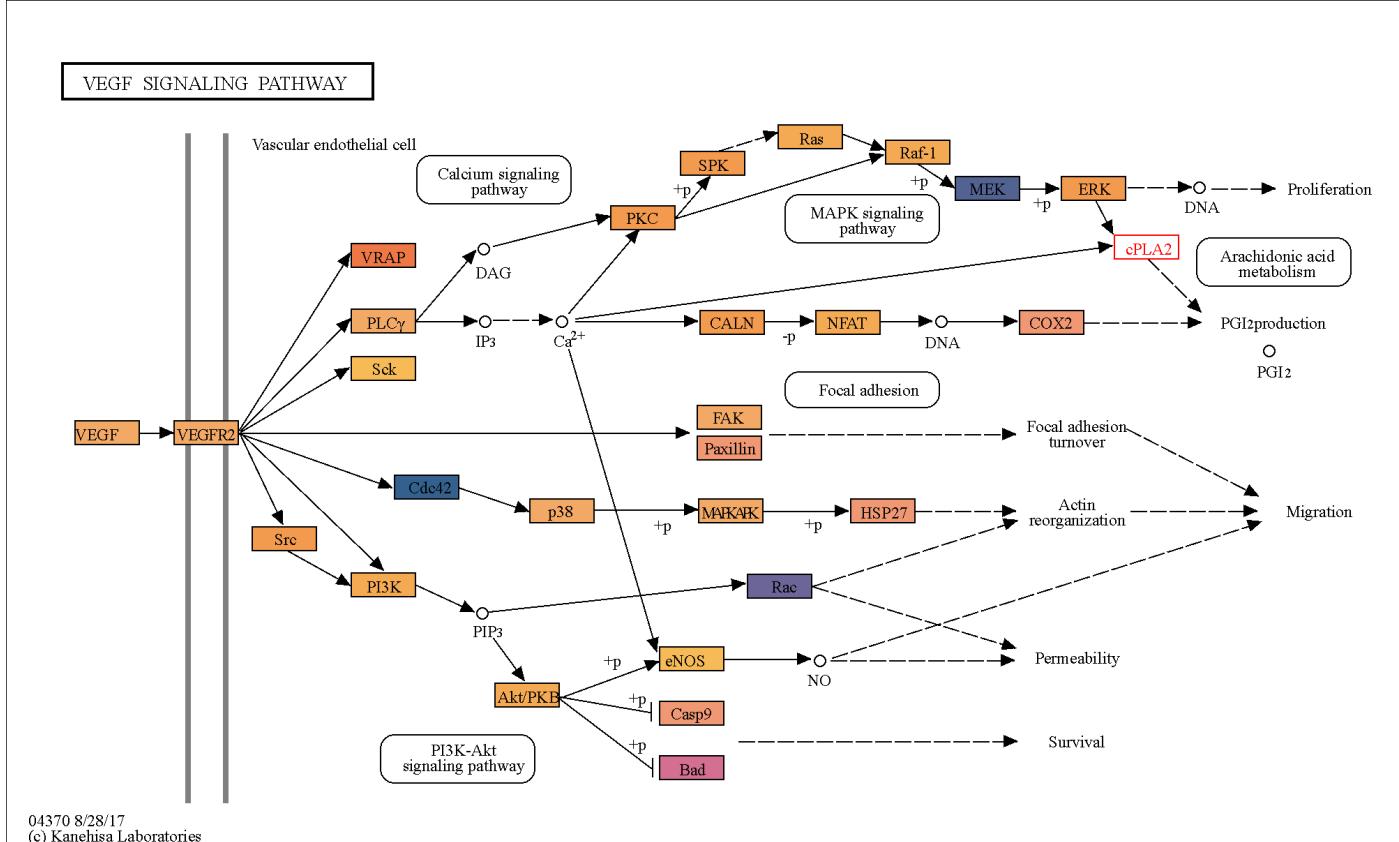
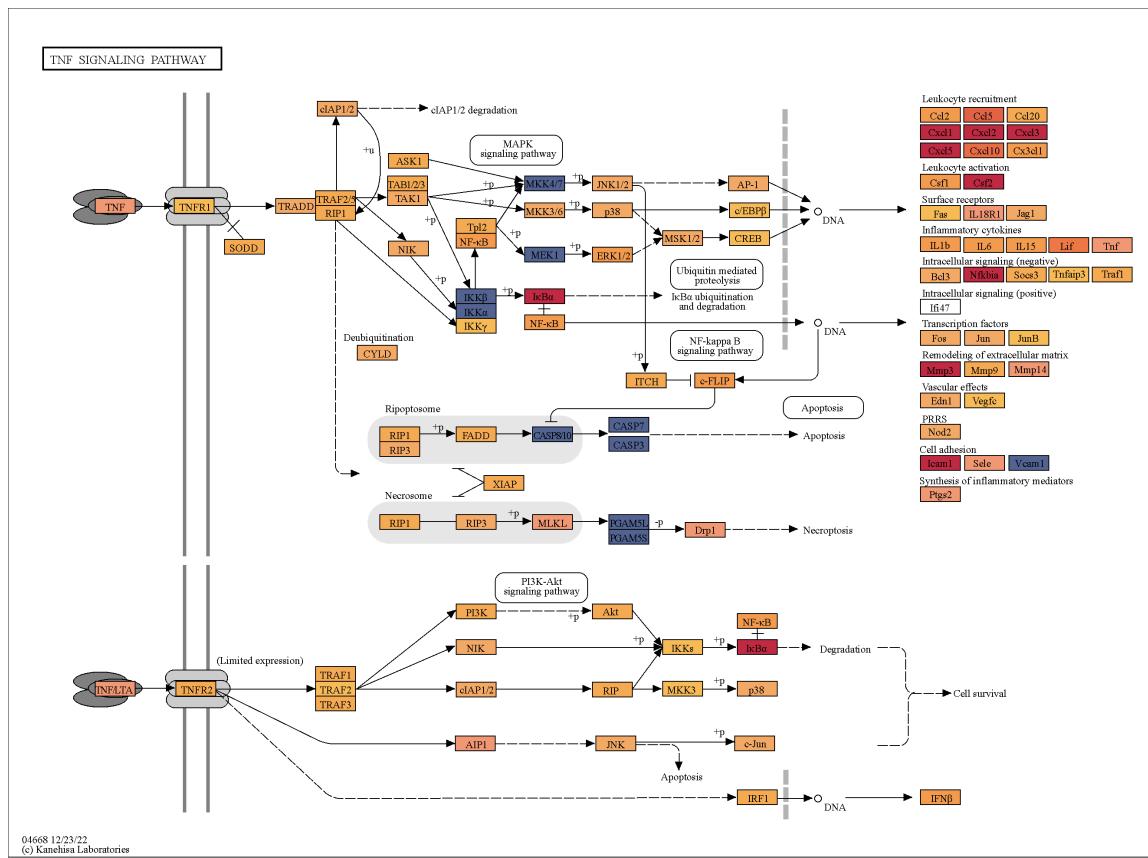


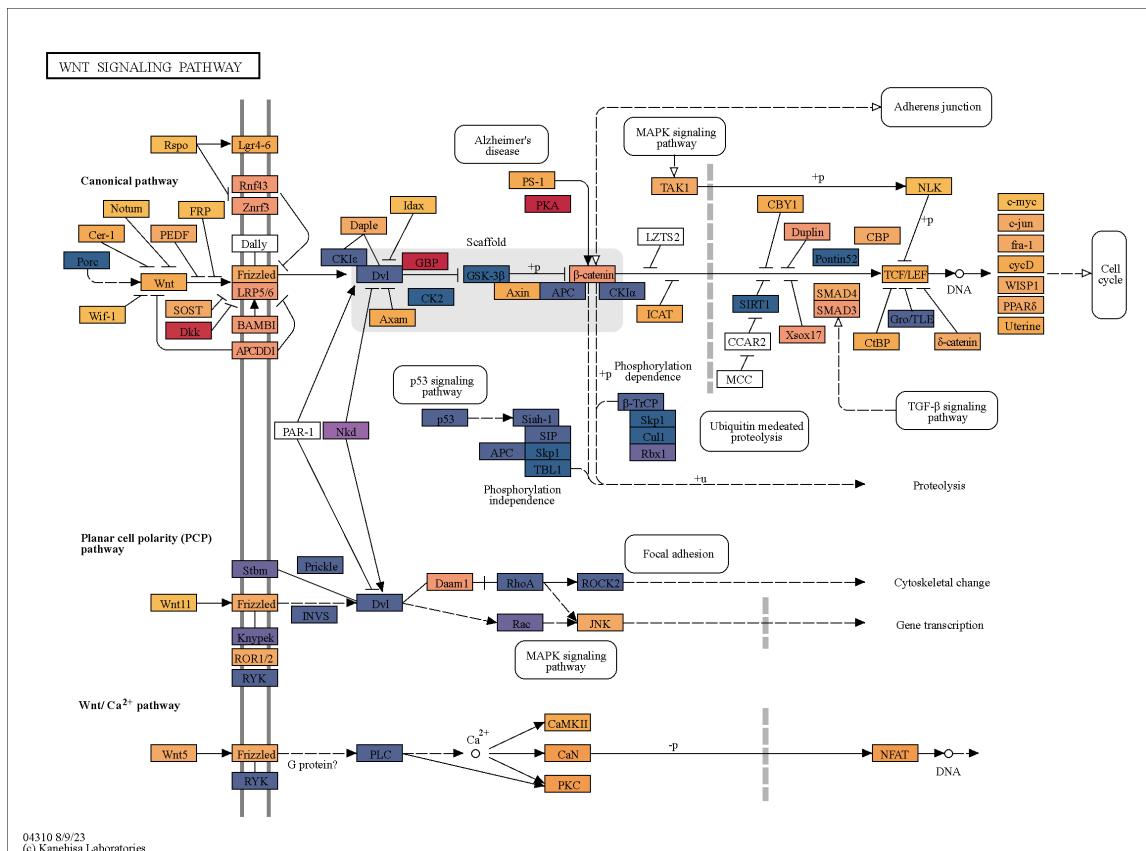








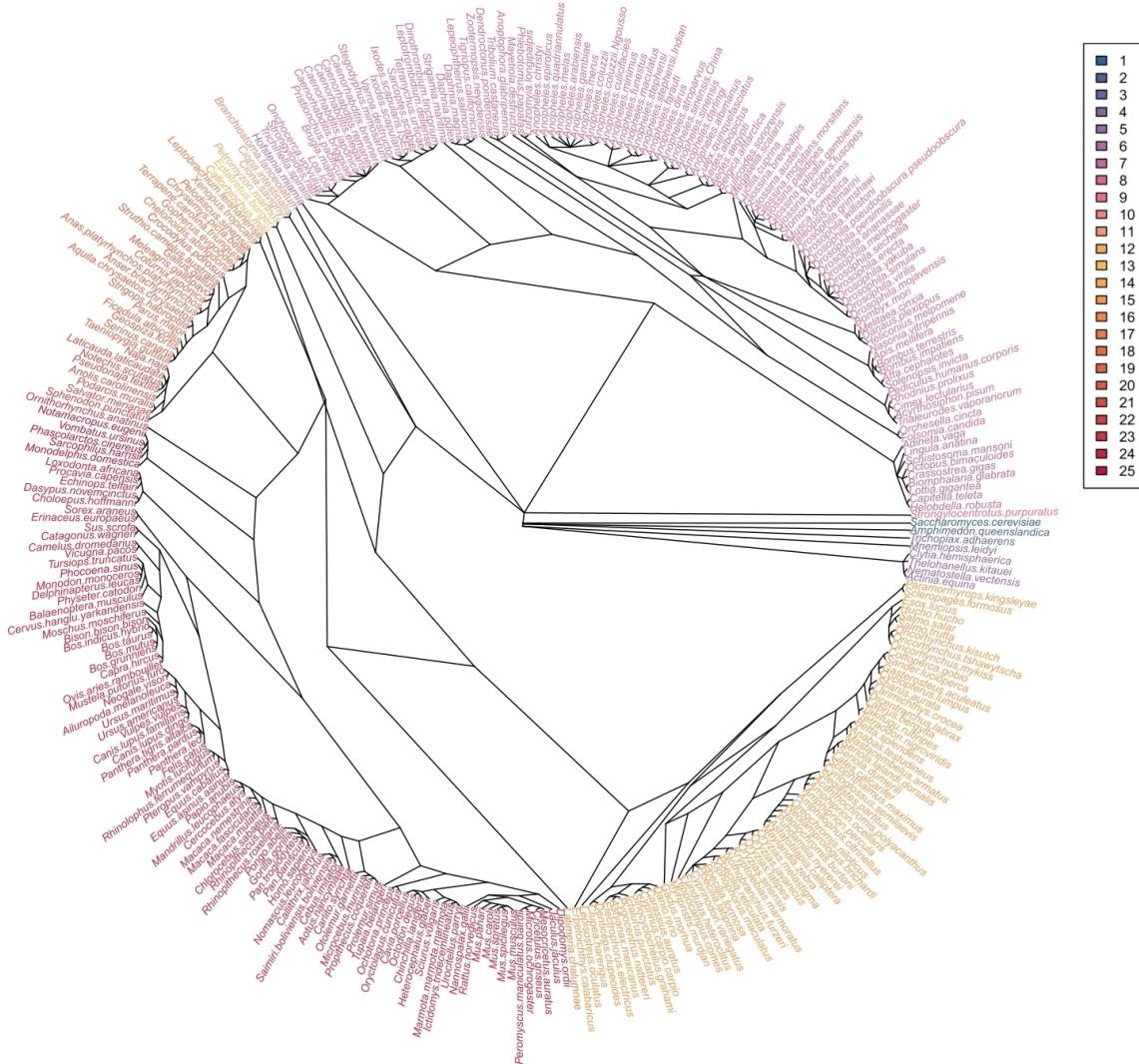




## 6.5 Article 1 - Suppl. Data 4

### Suppl. Data 4 - Animal tree of life and clades of study

Legend : Tree of life of the 315 species studied here, generated using the information available in Ensembl and Ensembl Metazoa, and with R's ape package (**paradis\_ape\_2019**). The tree is rooted to reflect the phylogeny of the Opisthokonta, and the branches are not to scale. The colors are those used in the figures of the article. Each clade is represented by one or more species in our database.



## Floriane PICOLO

### Étude de l'évolution des gènes codant les protéines de transduction des signaux intracellulaires chez les animaux

#### Résumé :

L'évolution des espèces est liée, entre autres, à l'évolution de leur génome. La fonction biologique des protéines codées par les gènes rend ces derniers sensibles aux influences d'autres gènes, en particulier ceux qui codent des protéines partenaires. Bien que nous ayons une compréhension croissante de l'organisation des gènes au sein des génomes, il est important de noter que les gènes peuvent mourir (disparaître), se transformer, ou naître, en particulier par le biais des duplications de gènes ou de génomes. Dans le contexte des voies de signalisation, les interactions entre les produits de gènes ont une importance particulière, car elles sont soigneusement régulées et ordonnées. L'objectif principal de cette thèse est de répondre à deux questions : premièrement, comment les voies de signalisation animales ont-elles évolué au fil du temps ? Deuxièmement, au sein d'un groupe d'espèces de vertébrés ayant subi trois ou quatre duplications complètes du génome (WGD), les gènes impliqués dans ces voies sont-ils restés en deux ou trois copies (duplicate ou triplicate), ou sont-ils revenus à une seule copie (singleton) ? Cette étude montre que parmi les 47 voies examinées, 24 d'entre elles ont évolué de manière ascendante, c'est-à-dire qu'elles ont émergé de la fin de la voie (du facteur de transcription) vers l'amont (les ligands et les récepteurs), tandis que 10 voies ont suivi un scénario contraire, se développant de l'amont vers l'aval. De plus, ces mêmes gènes sont restés généralement en deux copies chez les espèces ayant subi trois duplications complètes du génome, et sont même présents en triplicate ou plus chez les espèces ayant subi quatre duplications complètes, parmi le clade des téléostéens. Ces résultats suggèrent que l'évolution des voies de signalisation s'est faite de façon non aléatoire dans l'histoire évolutive des génomes animaux.

**Mots clés :** évolution, phylogénie, voie de transduction, duplication de génome, téléostéens

#### Abstract :

The evolution of species is linked, among other things, to the evolution of their genome. The biological function of proteins encoded by genes makes them susceptible to influences from other genes, particularly those that encode partner proteins. Although we have a growing understanding of the organization of genes within genomes, it is important to note that genes can be changed, lost, or born, particularly through gene or genome duplications. In the context of signaling pathways, interactions between gene products are of particular importance because they are carefully regulated and ordered. The main goal of this thesis was to answer two questions: first, how have animal signaling pathways evolved over time? Second, within a group of vertebrate species that have undergone three or four complete genome duplications (WGD), have the genes involved in these pathways remained in two or three copies (duplicate or triplicate), or have they returned to a single copy (singleton)? The results of this study show that among the 47 pathways examined, 24 of them evolved bottom-up, that is, they emerged from the end of the pathway (of the transcription factor) towards the upstream (ligands and receptors), while 10 pathways followed an opposite scenario, developing from upstream to downstream. Furthermore, these same genes generally remained in two copies in species having undergone three complete genome duplications, and are even present in triplicate or more in species having undergone four complete duplications, among the teleost clade. These results suggest that the evolution of signaling pathways occurred in a non-random manner in the evolutionary history of animal genomes.

**Keywords :** evolution, phylogeny, signaling transduction, whole genome duplication, teleosts