Compétition pour la transcription et évolution de l'expression génétique chez les diploïdes

Frédéric Fyon

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'Université de Montpellier

Préparée au sein de l'école doctorale GAIA
Et de l'unité de recherche UMR 5175 - CEFE

Spécialité : EERGP – Ecologie, Evolution, Ressources Génétiques, Paléobiologie

Présentée par Frédéric Fyon

Compétition pour la transcription et évolution de l'expression génétique chez les diploïdes

Soutenue le 13/12/2016 devant le jury composé de

Dr. Thomas Lenormand, DR CNRS, UMR 5175 - CEFE	Directeur de thèse
Dr. Sylvain Billiard, MCF, Université Lille 1	Rapporteur
Dr. Gabriel Marais, DR CNRS, UMR 5558 - LBBE	Rapporteur
Dr. Ophélie Ronce, DR CNRS, UMR 5554 - ISEM	Examinateur
Dr. Jan Engelstädter, Lecturer, University of Queensland	Examinateur

Cette thèse est dédiée aux collègues précaires,

Victimes inconnues de l'obsession financière,

Tiraillé-es tous les jours entre une rage amère

Et une passion scientifique nourricière.

Le monde de demain ne sera pas celui d'hier,

Rien ne nous oblige à rester dans la galère.

Sommaire

Introduction Générale

La régulation de l'expression des gènes est, depuis quelques années, au centre d'importantes recherches en biologie évolutive. Depuis les travaux précurseurs de King & Wilson (KING and WILSON 1975), il est admis que les séquences de régulation de l'expression génétique tiennent un rôle important dans l'évolution des organismes vivants. Pourtant, de nombreuses facettes des réseaux de régulation restent encore largement inexpliquées, et sujets d'intenses débats. Certains affirment, par exemple, que l'architecture complexe de ces réseaux est adaptative, alors que d'autres l'expliquent par des processus neutres. Ce débat entre adaptation et processus neutre se retrouve à tous les niveaux des réseaux de régulation.

En étudiant la littérature qui s'intéresse aux forces évolutives ayant influencé les réseaux de régulation, on remarque qu'il n'est jamais fait mention de sélection à l'échelle du gène, et non à l'échelle de l'individu. En effet, les mutations n'augmentent pas toutes en fréquence de façon neutre ou parce qu'elles améliorent le phénotype des individus. Certaines augmentent en fréquence simplement parce qu'elles ont un avantage de transmission héréditaire. Les éléments génétiques « égoïstes » en fournissent de nombreux exemples (BURT and TRIVERS 2009).

Dans cette thèse, nous nous sommes intéressés à la façon dont les séquences *cis*-régulatrices (voir définition plus bas) influent sur la sélection sur le gène régulé, et comment cette influence entraîne l'augmentation de fréquence de certaines séquences régulatrices sans que celles-ci apportent *in fine* de bénéfices aux individus.

Composition et fonctionnement des réseaux de régulation eucaryotes

Focalisation sur l'étape d'initiation de la transcription

La régulation de l'expression des gènes a lieu tout au long de la conversion de l'information génétique en protéines : initiation de la transcription, vitesse d'élongation, épissage alternatif, maturation, dégradation et exportation des ARNm hors du noyau cellulaire, traduction, sont toutes des étapes soumises à régulation (OGBOURNE and ANTALIS 1998; WRAY *et al.* 2003). Ici, nous nous intéresserons uniquement à la première étape, la régulation de l'initiation de la transcription, qui est bien connue. Cette étape semble être une étape majeure de régulation de l'expression (WRAY *et al.* 2003).

L'initiation de la transcription consiste en l'arrivée du complexe protéique ARN-polymérase II (LEE and YOUNG 2000), généralement au niveau d'une *TATA*-box (WRAY *et al.* 2003). La transcription démarre quelques bases plus près du gène, là où se trouve la partie active du complexe enzymatique (WRAY *et al.* 2003). Le site où démarre la transcription est appelé site d'initiation de la transcription (*Transcription Start Site*, ou *TSS*, en anglais).

Machinerie transcriptionnelle

La transcription requiert la co-occurence de nombreuses protéines autour des sites d'initiation de la transcription. L'ARN-polymérase II est l'enzyme chargée de la transcription : elle lit les bases d'un brin d'ADN, et synthétise un brin d'ARN. La reconnaissance et la fixation de l'ARN-polymerase II sur la *TATA*-box implique la présence à proximité de différentes protéines, appelées cofacteurs (LEE and YOUNG 2000). Ces cofacteurs sont généralement liés à des facteurs de transcription (*Transcription Factors*, ou *TF*, en anglais). Les facteurs de transcription sont des protéines qui possèdent, au moins, un domaine de fixation à l'ADN, et un domaine de fixation à d'autres protéines (WRAY *et al.* 2003). Le domaine de fixation à l'ADN est pourvu d'un site de reconnaissance de l'ADN, capable de reconnaître un certain nombre de séquences d'ADN proches selon différentes affinités (WRAY *et al.* 2003). La fixation des *TFs* sur ces séquences, puis la fixation des co-facteurs sur les *TFs* créé une signalisation élaborée à destination de l'ARN-polymerase II.

Les séquences régulatrices, en cis et en trans

Les séquences reconnues par les *TFs* sont appelées promoteurs, et sont de deux types. Les promoteurs centraux (*promoters* ou *core promoters* (BUTLER and KADONAGA 2002) en anglais) sont regroupés dans une centaine de paires de bases situées près des gènes, en amont (LEE and YOUNG 2000). Ces promoteurs centraux varient, mais certaines séquences, comme la *TATA-box,* se retrouvent identiques chez de nombreux gènes (WRAY *et al.* 2003). Ils permettent de fixer les composants du complexe de pré-initiation, composants incontournables pour fixer l'ARN-polymerase II (LEE and YOUNG 2000; WRAY *et al.* 2003).

Les promoteurs au sens large, aussi appelés amplificateurs (*enhancers* en anglais) peuvent être situés plus ou moins loin du gène (WRAY *et al.* 2003). Beaucoup plus variables, ces séquences attirent des facteurs de transcription chargés de réguler l'initiation de la transcription (SERFLING *et al.* 1985). Lorsque le promoteur est situé à une grande distance du gène, le facteur de transcription correspondant peut être malgré tout mis à proximité de la *TATA*-box par la création de boucles dans la configuration spatiale du brin d'ADN (WRAY *et al.* 2003; SEN and GROSSCHEDL 2010).

Généralement, on distingue *TFs* et co-facteurs d'un côté, et promoteurs (*enhancers* et *promoters*) de l'autre, sous les noms de *trans*- et *cis*-régulateurs. Chaque protéine de *TFs* ou de co-facteurs provient d'un gène sur un brin d'ADN, et est capable d'agir sur la transcription d'un gène situé sur n'importe quel brin d'ADN : on dit qu'ils régulent en *trans*. Au contraire, les promoteurs ne sont capables de réguler, en fixant tel ou tel *TF*, que le gène situé sur le même brin d'ADN, et non sur le brin homologue : on dit qu'ils régulent en *cis*.

Deux thèses sur la régulation

La base mécanistique de la régulation de l'initiation de la transcription n'est pas encore complètement résolue, et deux modèles s'affrontent (SEN and GROSSCHEDL 2010). Le modèle le plus courant affirme que les promoteurs permettent d'accélérer ou de ralentir l'initiation de la transcription. Cela signifie que la fixation de l'ARN-polymerase II peut être accélérée ou ralentie selon les *TFs* attirés par les promoteurs. Ainsi, un promoteur plus « fort » (accélère l'initiation de la transcription) permet à davantage d'ARN-polymerases II de se fixer sur le *TSS*.

Le second modèle est plus probabiliste. Il explique qu'il y a initiation de la transcription lorsque les *promoters* sont dans un état « activés ». Même sans *enhancers*, les *promoters* sont capables d'être activés. Les *enhancers*, et la cohorte de protéines qui va avec, ont ici pour rôle d'augmenter ou de diminuer le nombre de *promoters* activés au sein d'une population de cellules, c'est-à-dire d'augmenter la probabilité que les *promoters* soient activés (WALTERS *et al.* 1995; SEN and GROSSCHEDL 2010). Une fois les *promoters* activés, toutefois, la vitesse d'initiation ne varie pas.

Les deux modèles ont reçu des validations empiriques, et il semble prudent de dire que la régulation de l'expression peut se faire selon les deux mécanismes (SEN and GROSSCHEDL 2010).

La régulation de l'expression, un réseau

Les multiples acteurs de la régulation de l'initiation de la transcription ont la particularité d'interagir beaucoup et de façon très variable entre eux. Un facteur de transcription peut ainsi interagir avec plusieurs promoteurs. Inversement, un promoteur peut être reconnu par plusieurs facteurs de transcription (WRAY *et al.* 2003). Les mécanismes ontologiques requièrent des régulations en cascade : une protéine active un ou plusieurs gènes, dont les produits vont à leur tour reconnaître des séquences régulatrices et activer d'autres gènes. Le cocktail de facteurs de transcription et co-facteurs exprimés permettent ainsi de produire, au bon moment et au bon endroit, les protéines utiles au développement de l'organisme. Cette cascade s'organise comme un réseau, avec des motifs récurrents (MILO *et al.* 2002; ALON 2007). Parmi ces motifs, on peut notamment citer les boucles de rétroaction positives et négatives (BRANDMAN and MEYER 2008).

Une boucle de rétroaction positive suppose que le produit d'un gène (éventuellement le produit distant, après plusieurs activations de gènes en cascade) active en retour l'expression de ce gène. Ce motif est notamment utile pour construire un signal bimodal. Ainsi, lorsqu'un gène impliqué dans une telle boucle est éteint, pas ou peu du produit est synthétisé. Dès qu'un signal, même relativement faible, vient à activer un tant soit peu le gène, le surplus de produit active en retour davantage le gène, augmentant la production et ce jusqu'à ce que l'activation ait atteint son niveau maximal. Cela permet que le gène demeure activé même après que le signal ait disparu.

Une boucle de rétroaction négative suppose à l'inverse que le produit d'un gène (ou de la cascade) inhibe en retour sa propre expression. Ce motif est particulièrement utile pour assurer des niveaux d'expression stables. En effet, lorsque le niveau d'expression augmente, le surplus de produit inhibe davantage l'activation du gène, jusqu'à ce que le niveau d'expression soit revenu au niveau initial. De même, lorsque le niveau d'expression diminue, le gène est moins inhibé (du fait d'une quantité plus faible de produit), de sorte que le niveau d'expression ré-augmente jusqu'à son niveau initial. Les boucles de rétroaction négatives semblent être un mécanisme particulièrement efficace pour les gènes soumis à une sélection stabilisante importante sur leur niveau d'expression.

Surimpression de marques épigénétiques

Les interactions entre *cis*- et *trans*-régulateurs se font au sein d'un contexte régulateur plus large. Notamment, elles dépendent de l'accessibilité des promoteurs, c'est-à-dire de la conformation de la molécule d'ADN (WRAY *et al.* 2003). Cette conformation dépend notamment des promoteurs et des facteurs et co-facteurs qu'elles attirent (WRAY *et al.* 2003), mais pas uniquement. Les marques épigénétiques éventuellement présentes sur le double brin d'ADN ou sur les histones (protéines structurant l'ADN) affectent également sa conformation (BONIFER and COCKERILL 2011). Un chromosome entier peut être ainsi éteint via des marques épigénétiques, comme un chromosome X sur deux chez les mammifères femelles (MOREY and AVNER 2010). Des gènes peuvent également individuellement être éteints par des marques épigénétiques (ESTELLER 2007). Les marques épigénétiques sont très particulières : à la fois plastiques au cours de la vie des individus, ces marques peuvent être transmises lors de divisions cellulaires, et donc éventuellement transmises d'un individu à son descendant.

Vision d'ensemble du réseau d'initiation de la transcription

L'organisation de la régulation de l'initiation de la transcription peut être résumée comme sur la figure 1.

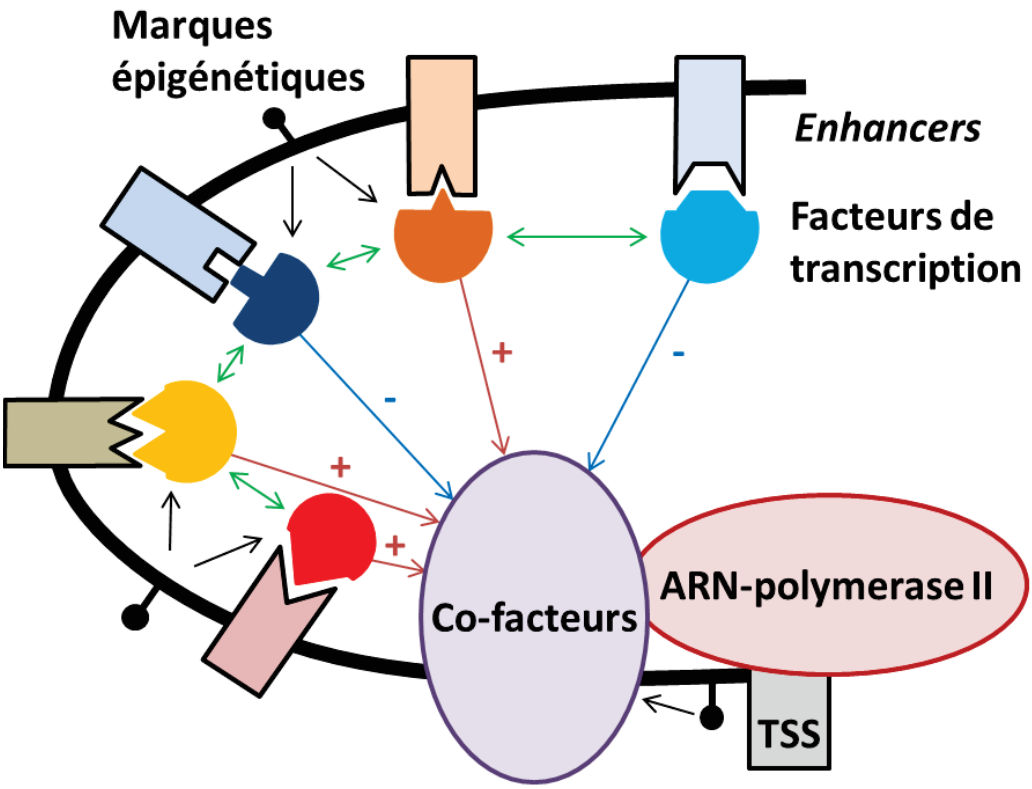


Figure 1. Représentation schématique des réseaux d'initiation de la transcription.

Les facteurs de transcription se fixent aux *enhancers* correspondants puis se lient à de nombreux co-facteurs, qui eux-mêmes permettent à l'ARN-polymerase II de se fixer au site d'initiation de la transcription (*TSS*). Les facteurs de transcription ont soit un effet activateur (**+**), soit un effet inhibiteur (**-**) sur l'agglomération des co-facteurs, et donc *in fine* sur l'initiation de la transcription. La fixation des facteurs de transcription dépend bien sûr des *enhancers* correspondants, mais pas seulement. L'accessibilité des sites de fixation peut être modulée par la présence d'autres facteurs de transcription sur des sites de fixation à proximité (flèches vertes sur fig. 1). La présence d'autres facteurs de transcription à proximité peut soit encombrer l'espace local, inhibant la fixation du facteur de transcription focal, soit reconfigurer l'espace local de sorte à rendre plus accessible le site de fixation du facteur de transcription focal. L'accessibilité des sites de fixation dépend également des

marques épigénétiques éventuellement présentes sur la molécule d'ADN ou sur les histones structurant cette molécule.

Evolution des réseaux de régulation

L'étude de l'évolution phénotypique par évolution de la régulation a commencé lorsque King & Wilson ont constaté que l'ampleur des variations des séquences d'acides aminés entre l'homme et le chimpanzé ne semblait pas suffisante pour expliquer l'ampleur des différences phénotypiques observées entre ces deux espèces (KING and WILSON 1975). Cela signifiait que la séquence d'une protéine ne fait pas tout : sa conformation précise, ainsi que son profil d'expression ont une influence non négligeable sur les phénotypes. Par profil d'expression, on entend : (1) le type cellulaire où la protéine est produite (différents tissus produisent différentes protéines), (2) le moment dans le cycle de vie d'un individu où la protéine est produite, (3) le niveau d'expression, c'est-à-dire la quantité de protéines produites. La conformation d'une protéine, comme son profil d'expression, dépendent du réseau de régulation de l'expression des gènes. Ainsi, il est fréquent d'observer des changements phénotypiques par changements de niveaux d'expression (RAYMOND *et al.* 1998; FEREA *et al.* 1999; COOPER 2003; ABZHANOV *et al.* 2004), par changements de temporalité d'expression (ABZHANOV *et al.* 2004), ou par changements de localisation d'expression (STERN 1998; WITTKOPP *et al.* 2002).

Séquences régulatrices et séquences codantes

Les organismes peuvent évoluer par des modifications soit dans les séquences des protéines, soit dans leurs profils d'expression. L'importance relative de ces deux types de mutations a fait l'objet de débats importants (Carroll 2005; Hoekstra and Coyne 2007; Wray 2007; Lynch and Wagner 2008).

Les approches dites d'"Evo-Devo" considèrent que les nouveautés adaptatives apparaissent le plus souvent dans les séquences *cis*-régulatrices (Carroll 2005; Wray 2007). D'abord, parce qu'elles supposent que l'effet de ces mutations sur les profils d'expression est co-dominant (Wray 2007). Comme chaque *cis*-régulateur contrôle un gène sur le même chromosome (en *cis*), et pas le gène homologue, certains en déduisent qu'une mutation *cis*-régulatrice ne modifie le profil d'expression que de la moitié des protéines correspondantes, d'où un effet co-dominant. Les mutations codantes étant généralement récessives, cette co-dominance rendrait les mutations *cis*-régulatrices plus visibles à la sélection que les mutations codantes, et donc plus promptes à répondre à la sélection. Ensuite, parce que les séquences *cis*-

régulatrices semblent organisées en modules distincts, composés chacun de plusieurs séquences de fixation de facteurs de trancription. Ces modules sont supposés correspondre chacun à un aspect particulier du profil d'expression, comme le niveau d'expression dans un type cellulaire particulier ou la réponse à un signal précis (Wray 2007; Wittkopp and Kalay 2012). Les mutations *cis*-régulatrices seraient donc très peu pléiotropes : des mutations bénéfiques peuvent être sélectionnées sans que d'éventuels effets pléiotropes coûteux viennent contrecarrer leur transmission.

Plusieurs études ont déjà montré des changements adaptatifs causés par des mutations dans les séquences *cis*-régulatrices (Carroll 2005; Wray 2007). Cependant, certains considèrent qu'affirmer que les mutations *cis*-régulatrices ont une place privilégiée dans les mécanismes d'adaptation est au mieux prématuré. Outre que de nombreux exemples d'adaptation par mutations codantes sont documentés (Hoekstra and Coyne 2007; Lynch and Wagner 2008), rien n'indique que les mutations codantes aient toutes de sérieuses conséquences pléiotropes. Par exemple, si la mutation apparait sur un gène récemment dupliqué, ses conséquences pleiotropes peuvent être minimisées par l'existence de la copie intacte du gène maintenant la fonction initiale(HOEKSTRA and COYNE 2007).

Les forces évolutives majeures s'appliquant aux réseaux de régulation

De nombreuses études se sont attachées à évaluer l'importance relative des différences forces évolutives à l'origine des modifications de séquences régulatrices (WHITEHEAD and CRAWFORD 2006a). Il en ressort que la sélection directionnelle et surtout la sélection stabilisante sont fréquentes (LUDWIG *et al.* 2000; DENVER *et al.* 2005; LEMOS *et al.* 2005; GILAD *et al.* 2006). Mais l'évolution des séquences régulatrices n'est pas uniquement dirigée par l'optimisation des profils d'expression.

De nombreuses mutations dans les séquences régulatrices sont neutres (KHAITOVICH *et al.* 2004; WHITEHEAD and CRAWFORD 2006b). Il peut s'agir soit de mutations qui ne modifient pas le réseau de régulation, soit de mutations qui le transforment en un réseau équivalent. En effet, plusieurs réseaux de régulation différents peuvent mener au même profil d'expression (WEIRAUCH and HUGHES 2010).

Des mutations peuvent augmenter sélectivement en fréquence sans apporter de bénéfice aux individus. Ces mutations augmentent en fréquence car elles obtiennent un bénéfice de transmission d'une génération à l'autre. Ces séquences sont appelés séquences égoïstes (BURT and TRIVERS 2009). Un exemple classique est les gènes de type « gamete killer », qui tuent tous les spermatozoïdes d'un individu ne portant pas ces gènes. Ils ont donc pour effet de réduire la charge en spermatozoïde de l'individu porteur, mais sont sûrs d'être transmis à la génération suivante même à l'état hétérozygote (BURT and TRIVERS 2009). Dans cette thèse, nous verrons que nous avons découvert un processus original qui conduit certains *cis*-régulateurs à augmenter en fréquence. Ces *cis*-régulateurs se répandent grâce à la formation d'associations génétiques (voir Encart 2) bénéfiques pour l'individu. Cependant, une fois fixé, ces *cis*-régulateurs n'apportent plus de bénéfices aux individus. Il s'agit bien d'un processus sélectif, dont la force motrice est l'avantage de cacher des mutations délétères, mais son caractère adaptatif n'est pas évident et est discuté.

Sujet d'étude : une nouvelle force sélective

L'évolution des séquences *cis*-régulatrices a souvent été étudié, mais rarement sous l'angle d'inégalités d'expression relative entre gènes homologues, que l'on appelle aussi expression allèle-spécifique (*Allele-Specific Expression*, ou *ASE*, en anglais). Les *ASE* sont très répandue dans les génomes (Lo *et al.* 2003). En effet, puisque chaque séquence *cis*-régulatrice contrôle l'expression d'un gène sur le même chromosome (en *cis*), un polymorphisme de ces séquences peut générer un patron d'*ASE*, ou une copie d'un gène est plus exprimée que la copie homologue. Cela modifie le phénotype des individus, principalement déterminé par la copie la plus exprimée. Ainsi, le polymorphisme *cis*-régulateur est susceptible d'influer sur les processus de sélection au niveau du gène en modifiant les coefficients de dominance des mutations sur ce gène (voir Encart 1). C'est cette influence, et ses conséquences sélectives pour les *cis*-régulateurs, qui forment le cœur de cette thèse.

Encart 1 - La Dominance

Soit un trait phénotypique *z*, dépendant d'un gène **A**. L'allèle sauvage est noté *A*, l'allèle mutant *a*. Alors, la dominance *h* de la mutation *a* sur le trait *z* se calcule comme :

$$h = \frac{z_{Aa} - z_{AA}}{z_{aa} - z_{AA}}$$

Plus une mutation est dominante (*h* tend vers 1), plus la valeur du trait chez l'hétérozygote est proche de la valeur du trait chez les individus homozygotes pour la mutation (z_{Aa} proche de z_{aa}). A l'inverse, plus une mutation est récessive, (*h* tend vers 0), plus le trait chez l'hétérozygote ressemble au trait chez les homozygotes pour l'allèle sauvage (z_{Aa} proche de z_{AA}). La dominance n'est pas une caractéristique de la mutation elle-même, mais dépend du trait considéré, du contexte génétique (par exemple les autres allèles au locus considéré), et du contexte environnemental.

Plusieurs théories ont été développées pour expliquer la dominance des mutations, en particulier la récessivité (faible coefficient de dominance) des mutations délétères. Une première théorie, basée sur des modèles d'évolution de gènes "modifieur" de dominance et lancée par les travaux précurseurs de Fisher (Fisher 1928), affirme que la dominance évolue pour minimiser l'effet des mutations délétères. Cette théorie souffre de plusieurs défauts.

Notamment, elle repose sur une pression de sélection faible, de l'ordre du taux de mutation. Une telle sélection faible peut être facilement contrecarrée, soit par la dérive génétique en petite population, soit par d'éventuels effets pléiotropes du gène modifieur (WRIGHT 1934a). Cette théorie a également été contredite par plusieurs observations empiriques (CHARLESWORTH 1979; ORR 1991).

La théorie du contrôle métabolique (WRIGHT 1934b; KACSER and BURNS 1981), affirme elle que la dominance des mutations est le résultat des mécanismes métaboliques, qui associent en réseaux plusieurs enzymes codées par des gènes différents. Ces enzymes partagent le contrôle des propriétés du réseau (en particulier le flux métabolique), si bien que la modification de l'activité d'une seule de ces protéines (par mutation) n'impacte que très peu le flux métabolique global, et seulement si cette activité est fortement réduite (i.e. chez l'homozygote muté et non chez l'hétérozygote). La récessivité des mutations délétères découlerait directement du fonctionnement métabolique des organismes. Cette théorie, prédominante à l'heure actuelle, pose également plusieurs problèmes. Elle n'est vraiment applicable qu'aux traits métaboliques. Elle suppose que la fitness est une fonction linéaire croissante du flux métabolique, ce qui exclut tout optimum intermédiaire (HURST and RANDERSON 2000) dû à d'éventuels *trade-offs* et ne permet pas d'expliquer les cas de super-dominance. La théorie du contrôle métabolique a également été contredite par des observations empiriques (SZAFRANIEC *et al.* 2003).

Dans les modèles que nous développons dans cette thèse, nous considérons, pour fixer les idées, que les mutations délétères sont partiellement récessives, avec $h = 0,25$ comme observé en moyenne (MANNA *et al.* 2011). Nous incorporons un locus *cis*-régulateur qui peut être vu comme un modifieur d'expression relative. Un gène associé (en *cis*) à un promoteur fort sera beaucoup exprimé. S'il fait face à une copie homologue associée à un promoteur faible, il sera à l'origine de l'essentiel de la production de la protéine correspondante. On suppose que cela influe sur le coefficient de dominance des mutations portées par le gène : une mutation sera beaucoup plus dominante si elle se retrouve dans l'essentiel des protéines produites que si elle n'impacte qu'une minorité des protéines exprimées. Les *cis*-régulateurs peuvent être considérés comme une forme originale de modifieur de dominance.

ER process (CHAPITRE 1)

Nous avons tout d'abord considéré un modèle multi-locus simple. Ce modèle comporte un locus gène, soumis à un régime de mutations délétères récurrentes. Il comporte également un locus promoteur (*cis*-régulateur), soumis à un régime de mutation impactant leur « force », c'est-à-dire leur capacité à activer l'initiation de la transcription. Plus un promoteur est fort, plus la copie du gène correspondante (i.e. en *cis*) est exprimée.

L'étude de ce modèle nous a permis de montrer que l'évolution du locus promoteur est soumise à deux effets sélectifs. On appelle le premier « effet de masquage ». Il est dû au fait que les mutations délétères, étant récessives, sont partiellement masquées à l'état hétérozygote. Lorsque l'on fait varier les expressions relatives des deux copies du gène chez l'hétérozygote, cela fait varier la dominance de l'allèle délétère. S'il n'y a pas d'associations génétiques (voir Encart 2), cette variation tend à augmenter la dominance moyenne, c'est-à-dire à « démasquer » les allèles délétères. Le polymorphisme au locus promoteur est ainsi contre-sélectionné : toute mutation au locus promoteur tend à disparaître rapidement parce qu'en moyenne elle démasque les mutations délétères. Le deuxième effet sélectif est appelé « effet de purge ». Il est dû au fait que les promoteurs forts ont tendance à purger les gènes associés en *cis* des mutations délétères. En effet, chez les individus double hétérozygotes (hétérozygotes aux deux locus gène et promoteur), l'allèle du gène le plus exprimé est l'allèle situé sur le même chromosome que le promoteur le plus fort. Cet allèle plus exprimé dicte davantage le phénotype de l'individu que l'allèle homologue. Il est donc davantage exposé à la sélection, et est mieux purgé des mutations délétères. Des associations génétiques (voir Encart 2) entre promoteurs forts et allèles viables d'un côté, et promoteurs faibles et allèles délétères de l'autre apparaissent, parce qu'elles permettent de mieux cacher les allèles délétères : ces associations sont favorisées parce que le coefficient de dominance moyen des allèles délétères diminue. Se retrouvant ainsi dans un contexte génétique favorable (associés en *cis* à des allèles viables), les promoteurs forts sont sélectivement favorisés, et augmentent en fréquence.

Le processus résulte en une escalade de la force des promoteurs jusqu'à ce qu'il n'existe plus de mutations capables d'augmenter encore davantage leur force. C'est pourquoi nous

l'avons nommé « processus de fuite en avant des promoteurs » (*Enhancer Runaway process*, ou *ER process*, en anglais).

Ce processus repose sur les associations génétiques entre les promoteurs et le gène. Ces associations sont moins fortes quand le taux de recombinaison entre le promoteur et le gène augmente (voir Encart 2). Si ce taux est trop élevé, les associations entre promoteurs forts et allèles viables sont régulièrement cassées, ce qui réduit considérablement l'effet de purge. Ne reste alors plus que l'effet de masquage, de sorte que les promoteurs mutants, même plus forts, sont contre-sélectionnés parce que la dominance moyenne des allèles délétères augmente. Au contraire, si le taux de recombinaison est suffisamment faible, l'association génétique entre promoteurs forts et allèles viables reste forte, la dominance moyenne des mutations délétères est faible et les promoteurs forts sont sélectionnés.

Nous prédisons donc que, pour chaque gène, il existe une fenêtre, à proximité du gène, où l'*ER process* peut avoir lieu. Au-delà de cette fenêtre, le promoteur est trop éloigné physiquement, le taux de recombinaison avec le gène est trop élevé pour que l'*ER process* puisse avoir lieu.

Ce processus est original dans la mesure où il comporte à la fois une dimension adaptative et non-adaptative pour les individus. Non-adaptative : le promoteur fort augmente en fréquence parce qu'il est transitoirement associé à un contexte génétique favorable, mais une fois fixé il n'apporte aucun avantage sélectif aux individus. Adaptative : le promoteur fort est associé à un contexte génétique favorable parce que cela permet de cacher les mutations délétères, associées à des promoteurs plus faibles, ce qui est sélectivement favorable pour les individus.

Encart 2 – Déséquilibre de liaison et recombinaison

Un déséquilibre de liaison désigne une association génétique préférentielle. Imaginons deux locus, **A** et **B**, ayant chacun deux allèles, *A* et *a*, *B* et *b*. Sur un chromosome, 4 haplotypes sont possibles : *AB*, *Ab*, *aB*, *ab*. Si l'on note p_A et p_B les fréquences de *A* et *B* dans la population, on s'attend à ce que les fréquences des 4 haplotypes soient : $p(AB) = p_A\ p_B$; $p(Ab) = p_A\ (1 - p_B)$; $p(aB) = (1 - p_A)\ p_B$; $p(ab) = (1 - p_A)(1 - p_B)$.

Cependant, plusieurs processus peuvent faire en sorte que les fréquences effectives des quatre haplotypes soient différentes de celles-ci. Lorsque c'est le cas on dit qu'il y a, entre les deux locus, des associations préférentielles d'allèles, c'est-à-dire, un déséquilibre de liaison.

Un déséquilibre de liaison peut être créé par la sélection lorsque l'effet sélectif du double mutant diffère du produit des effets de chacun des allèles (i.e. en présence d'épistasie). Ainsi, s'il y a un avantage particulier à avoir l'association *AB*, cette association génétique sera statistiquement sur-représentée après chaque étape de sélection. La migration entre populations présentant des fréquences alléliques différentes à plusieurs loci peut également générer un déséquilibre de liaison. Le déséquilibre de liaison peut varier stochastiquement. Il peut varier par dérive génétique : une association génétique peut résulter d'un effet d'échantillonnage en population finie. Il peut aussi varier par le hasard des mutations : une mutation apparaît aléatoirement dans un certain contexte génétique, et est donc initialement nécessairement associée à un contexte génétique particulier. Lorsque cette variance stochastique (par mutation ou dérive) est associée à de la sélection directionnelle, un déséquilibre de liaison négatif tend à être produit (effet Hill-Robertson).

Les déséquilibres de liaison, positifs ou négatifs, sont réduits par l'effet de la recombinaison. La recombinaison est un événement de brassage génétique qui a éventuellement lieu surtout pendant la méiose. Elle se déroule en prophase de 1ère division, lorsque les chromosomes homologues s'apparient. Cet appariement peut se résoudre sans échange génétique. Lorsqu'il y a échange génétique, il peut y avoir soit un événement de conversion génique, qui transforme une des copies d'un gène en la copie homologue *(AB / ab* devient *AB / aB* par exemple*)*, soit un événement de recombinaison, qui échange deux copies homologues entre elles (*AB / ab* devient *Ab / aB*). Du fait de la recombinaison, des associations sur-représentées tendent à être transformées en associations sous-représentées, ce qui réduit le déséquilibre de liaison. Ainsi, si l'on note r le taux de recombinaison entre deux gènes, D et D' les déséquilibres de liaison à t et $t + 1$ et si l'on suppose que le déséquilibre de liaison n'est entretenu par aucun processus, la recombinaison réduit le déséquilibre de liaison de sorte que $D' = (1 - r)D$.

Originalité du modèle (Chapitre I)

Le modèle étudié ici ressemble à plusieurs égards à des modèles qui existent déjà. Il se rapproche par exemple fortement des modèles d'évolution de la ploïdie (OTTO and GOLDSTEIN 1992; CAILLEAU *et al.* 2010). En effet, la variabilité de force des promoteurs peut être vue comme une variabilité graduelle de ploïdie: un gène qui possède un promoteur très fort sur un chromosome, et un promoteur très faible sur le chromosome homologue a, de fait, une expression quasi-haploïde. Les promoteurs pourraient être considérés comme des modifieurs de ploïdie. Dans ces modèles comme dans le nôtre, l'haploïdie entraîne une forte purge des mutations délétères qui bénéficie au modifieur de ploïdie si celui-ci est suffisamment proche du gène. Notre modèle s'éloigne toutefois des modèles de ploïdie en ceci que, une fois qu'un promoteur fort a envahi, a transitoirement modifié la ploïdie et s'est fixé dans une population, l'expression diploïde est restaurée (un promoteur fort mène à une expression plus haploïde uniquement s'il est associé à un promoteur faible). Notre modèle est également différent dans la mesure où le fardeau de mutations ne varie pas (toujours *2u* si *u* est le taux de mutation), alors qu'il varie entre *u* et *2u* dans les modèles de ploïdie.

Notre modèle se rapproche également des modèles d'évolution de la dominance (FISHER 1928; WAGNER and BÜRGER 1985; OTTO and YONG 2002; PROULX and PHILLIPS 2005) dans la mesure où, comme nous l'avons dit, la variabilité des forces des promoteurs modifient les rapports de dominance entre copies homologues du gène. Les promoteurs peuvent être considérés comme des modifieurs de dominance. Comme dans ces modèles, l'invasion d'un modifieur est liée à une baisse du coefficient de dominance des allèles délétères du gène. Toutefois, la même remarque que précédemment peut être faite : une fois que le promoteur mutant a envahi, la dominance initiale est restaurée. La dominance n'évolue pas, elle est transitoirement modifiée. Une autre différence importante est dans la nature du modifieur de dominance. Dans les modèles classiques, ce modifieur agit en *trans*, de sorte qu'à un génotype correspond un coefficient de dominance de l'allèle délétère donné. Dans notre modèle, le modifieur agit en *cis*. A chaque génotype correspond alors deux coefficients de dominance possibles, selon que l'allèle délétère est associé au promoteur fort ou au promoteur faible.

ER process et coévolution (CHAPITRE II)

L'escalade de la force des promoteurs peut avoir un effet négatif pour les individus en termes de niveau d'expression si ceux-ci ne sont pas contrôlés par une boucle de rétroaction négative assurant des niveaux d'expression optimaux. Au-delà d'un certain seuil, en effet, il devient délétère de produire trop de protéines. Dans un modèle comportant seulement deux locus, pas de boucle de rétroaction et avec de la sélection stabilisante sur les niveaux d'expression, l'augmentation des forces des promoteurs n'a pas lieu.

Toutefois, un tel modèle ne prend pas en compte une dimension majeure des réseaux de régulation, déjà mentionnée plus haut : l'existence de réseaux équivalents, le fait que plusieurs réseaux différents peuvent mener aux mêmes profils d'expression, et notamment aux mêmes niveaux d'expression (WEIRAUCH and HUGHES 2010). Cela signifie notamment qu'une mutation perturbant un niveau d'expression peut être compensée par une autre mutation dans un autre régulateur.

En construisant des modèles à trois locus, un gène et deux régulateurs (*cis* ou *trans*), nous montrons que l'augmentation des forces d'un régulateur en *cis* peut être compensée par l'évolution de *cis*- ou de *trans*-régulateurs restaurant un niveau d'expression optimal. La fuite en avant des promoteurs n'est alors pas stoppée, mais simplement ralentie par la sélection stabilisante sur les niveaux d'expression. Cette problématique est abordée tout d'abord dans le Chapitre I, avant d'être approfondie dans le Chapitre II avec une comparaison de la rapidité de coévolution en fonction de la nature de la séquence régulatrice qui coévolue avec le promoteur.

Les promoteurs, des séquences égoïstes ? (Chapitre II)

L'invasion d'un promoteur mutant fort pose de nombreuses questions en termes de niveaux de sélection. Ce promoteur envahit grâce à une association génétique favorable pour le promoteur, mais aussi pour les individus porteurs (allèles délétères partiellement cachés). La sélection semble donc se dérouler au niveau de l'individu. Pourtant, une fois le promoteur mutant fixé, non seulement les allèles délétères ne sont plus cachés, mais en plus les niveaux d'expression du gène ont augmenté, ce qui peut être délétère. De ce point de vue, l'invasion des promoteurs semblent plutôt égoïste, et la sélection semble agir au niveau du gène. On

s'attend à ce que cela aboutisse à un conflit génétique : un suppresseur envahit qui empêche les séquences égoïstes de se répandre, ce qui est bénéfique pour le suppresseur et l'individu puisque cela réduit les coûts liés à la propagation des séquences égoïstes.

Ce qu'il se passe dans notre modèle n'est pas un conflit génétique aussi simple. Pour réduire les coûts de surexpression liés aux promoteurs forts qui envahissent, une coévolution peut avoir lieu au sein du réseau de régulation, qui maintient des niveaux d'expression optimaux malgré des promoteurs de plus en plus fort. Si cela permet effectivement d'annuler les coûts de surexpression, cette coévolution a l'effet original de ne pas empêcher l'invasion de nouveaux promoteurs « égoïstes », mais au contraire de la rendre possible. Il ne s'agit donc pas là d'un conflit génétique au sens classique du terme, mais d'un processus continu de compensation des coûts induits par l'invasion des séquences égoïstes.

Influence des modes de reproduction (Chapitre III)

L'*ER process* dépend de l'hétérozygotie de la population, puisque c'est chez les double hétérozygotes (hétérozygotes aux locus gène et promoteur) que des différences de fitness amène à une sélection indirecte pour les promoteurs plus forts. Or, la fréquence d'hétérozygotes dans la population dépend elle-même, entre autres, du mode de reproduction. Ainsi, une population dont les individus se reproduisent par autofécondation aura une hétérozygotie inférieure à une population dont les individus se croisent en panmixie. L'*ER process* dépend également beaucoup du taux de recombinaison entre le gène et son promoteur : il a lieu à des faibles taux de recombinaison, lorsque les associations génétiques à la base du processus sont conservées. Le taux de recombinaison « efficace » tend à diminuer en autofécondation, car il y a plus d'homozygotes, chez lesquels la recombinaison n'a aucun effet. L'autofécondation tend donc à diminuer la fréquence d'hétérozygotes, et à diminuer la recombinaison efficace, deux effets aux conséquences a priori opposées en termes d'intensité de l'*ER process*.

Dans le Chapitre I, nous montrons que l'autofécondation tend à ralentir l'*ER process*. Dans le Chapitre III, nous étendons cette étude aux différents cas d'asexualités (reproduction mitotique et diverses formes de parthénogénèses) afin de savoir si les effets sont les mêmes quel que soit le mode d'asexualité. Certaines formes d'asexualité, comme la reproduction mitotique, présentent qui plus est un isolement génétique des chromosomes homologues.

Cet isolement entraîne de forts taux d'hétérozygotie, favorisant a priori l'*ER process*. Cependant, puisque les chromosomes sont isolés, les promoteurs mutants sont restreints au contexte génétique qui les a vu apparaître, et ne peuvent donc pas envahir toute la population. Il n'est donc pas évident de prédire ce qu'il peut se passer dans ces cas où les chromosomes homologues sont génétiquement isolés.

Dans les cas où les chromosomes ne sont pas isolés, l'*ER process* peut avoir lieu. La vitesse de l'*ER process* est positivement corrélée avec l'hétérozygotie de la population, elle-même positivement corrélée au taux d'allogamie. En effet, les modes de reproduction qui s'éloignent de l'allogamie mais pour lesquelles les chromosomes homologues ne sont pas isolés (autogamie et certaines formes de parthénogénèse) tendent à produire des homozygotes. Dans ces cas-là, on constate que la vitesse de l'*ER process* diminue à mesure que les taux d'allogamie diminuent, jusqu'à être nulle lorsqu'il n'y a plus du tout d'allogamie.

Dans les cas où les chromosomes sont isolés (reproduction mitotique, certains cas de parthénogénèse) l'évolution de la force des promoteurs suit un patron complètement différent. Dans ces cas-là, dans chaque lignée, une copie du locus promoteur devient de plus en plus forte, tandis que la copie homologue devient de plus en plus faible. Progressivement, l'expression du gène devient haploïde. En parallèle, la copie « cachée » du gène (celle avec le promoteur faible) accumule des mutations délétères puisqu'elle n'est pas ou peu soumise à la sélection purifiante, tandis que la copie exprimée (celle avec le promoteur fort) est purgée des mutations délétères. Nous appelons processus de divergence des promoteurs (*Enhancer Divergence process*, ou *ED process* en anglais) ce second processus, qui dérive de la même sélection indirecte sur les séquences *cis*-régulatrices que l'*ER process*.

ER et *ED process* peuvent être concomittants, à de forts taux d'isolements génétiques entre chromosomes homologues. Au fur et à mesure que cette isolement diminue, il y a de moins en moins divergence, et de plus en plus escalade, mais ces processus se chevauchent, pouvant amener une certaine diversité de patrons évolutifs selon les cas.

Le cas particulier des chromosomes sexuels (Chapitre IV)

Le facteur décisif de la transition entre *ER* et *ED process* est l'isolement génétique des chromosomes homologues. Cet isolement fait que la propagation de toute mutation est limitée au contexte génétique dans lequel elle est apparue. En d'autres termes, les lignées clonales sont hétérozygotes à la plupart des loci : il y a donc forcément une copie de chaque gène avec un promoteur plus fort que la copie homologue.

Les chromosomes sexuels ont une transmission assez proche de celles des lignées clonales. Dans le sexe hétérozygote (XY ou WZ), l'absence de recombinaison entre les chromosomes sexuels fait que ceux-ci sont hérités comme le matériel génétique d'organismes haploïdes asexués. L'isolement génétique des chromosomes sexuels chez le sexe hétérozygote est donc similaire à celui des lignées clonales étudiées dans le Chapitre III, à la différence près que les chromosomes X se transmettent, chez les femelles, de manière sexuée. Nous nous sommes donc intéressés aux modalités d'évolution des promoteurs dans le cas où les locus gène et promoteur sont présents sur des jeunes chromosomes sexuels, qui ont cessé de recombiner mais dont le chromosome hémizygote (Y ou Z) n'a pas encore dégénéré.

Dans un tel cas de figure, on retrouve effectivement un processus de divergence des promoteurs entre chromosomes X et Y (Z et W). De façon intéressante, ce processus mime parfaitement le processus supposé d'évolution des chromosomes sexuels (Orr and Kim 1998; Ercan 2015): des mutations délétères s'accumulent sur le chromosome Y (W), l'expression du chromosome Y (W) diminue jusqu'à extinction, l'expression des chromosomes X (Z) augmente chez les mâles (femelles) et éventuellement des compensations de dosage ont lieu sur les chromosomes X (Z) des femelles (mâles). On constate que ce processus est surtout efficace pour des grandes populations (car processus de sélection faible), alors que les effets d'interférence sélective classiquement considérés semblent plutôt efficaces en petites populations. Ces résultats nous amènent à penser que l'*ED process* a pu participer aux multiples évolutions convergentes des chromosomes sexuels.

Confirmation empirique du ER process *(Chapitre V)*

L'essentiel de cette thèse a porté sur une approche théorique des causes et conséquences évolutives de la sélection indirecte étudiée sur les réseaux de régulation. Toutefois, afin de tester la théorie développée ici, nous voulions apporter une première confirmation empirique du *ER process*. Pour ce faire, nous avons d'abord défini un signal qui ne peut être expliquée que par l'*ER process*. L'*ER process* consiste en une fuite en avant, un *runaway* de promoteurs de plus en plus forts qui envahissent la population. Il a donc pour effet d'accélérer, à travers une sélection positive, la divergence des séquences régulatrices, en particulier celles proches du gène. Le signal que nous étudions consiste donc en une divergence des séquences régulatrices proches du gène supérieure à celle des séquences plus éloignées (la recombinaison ralentit l'*ER process*), et supérieure à la divergence neutre. Pour tester ce signal, nous avons utilisé des données de divergence entre la souris (*Mus musculus*) et le rat (*Rattus norvegicus*) tirées de la banque de données Ensembl. Les résultats obtenus montrent que le signal, bien que faible, est détectable. Il s'agit là d'une première confirmation empirique du *ER process*.

Article published in Plos Genetics

Available at http://dx.doi.org/10.1371/journal.pgen.1005665

Chapitre I

Enhancer Runaway and the Evolution of Diploid Gene Expression

Authors : Frédéric Fyon, Thomas Lenormand

UMR 5175 CEFE, CNRS - Université Montpellier - Université P. Valéry - EPHE, 1919 route de Mende 34293 Montpellier Cedex 5, France

Abstract

Evidence is mounting that the evolution of gene expression plays a major role in adaptation and speciation. Understanding the evolution of gene regulatory regions is indeed an essential step in linking genotypes and phenotypes, and in understanding the molecular mechanisms underlying evolutionary change. The common view is that expression traits (protein folding, expression timing, tissue localization and concentration) are under natural selection at the individual level. Here, we use a theoretical approach to show that, in addition, in diploid organisms, enhancer strength (i.e., the ability of enhancers to activate transcription) may increase in a runaway process due to competition for expression between homologous enhancer alleles. These alleles may be viewed as self-promoting genetic elements, as they spread without conferring a benefit at the individual level. They gain a selective advantage by getting associated to better genetic backgrounds: deleterious mutations are more efficiently purged when linked to stronger enhancers. This process, which has been entirely overlooked so far, may help understand the observed overrepresentation of cis-acting regulatory changes in between-species phenotypic differences, and sheds a new light on investigating the contribution of gene expression evolution to adaptation.

Author's Summary

With the advent of new sequencing technologies, the evolution of gene expression regulation is becoming a subject of intensive research. In this paper, we report an entirely new phenomenon acting on the evolution of gene regulatory sequences. We show that in a small genomic region around genes there is a selection pressure to increase expression, such that stronger enhancers are favored. This leads to an open-ended escalation of enhancer strength. This outcome is not a particular case and we expect it to occur for all genes in nearly all eukaryotic diploid organisms. We also show that this escalation is not stopped by stabilizing selection on expression profiles. Indeed, regulators may coevolve to maintain optimal phenotypes despite the enhancer strength escalation. This widespread phenomenon can significantly shift our understanding of gene regulatory regions and opens a wide array of possible tests.

Introduction

The evolution of gene expression has become a subject of intensive research in the last years, sparking debates upon its role in adaptive evolution and speciation (1–6). Clearly, protein folding, expression levels, timing and tissue localization of expression are important regulatory traits under selection as they are essential steps in the genotype-to-phenotype map.

Gene expression is regulated at each step along the pathway from DNA to protein. Among them, transcription initiation is a crucial step responsible for a large proportion of the variation in expression profiles. Here we will focus on regulatory regions controlling this transcription initiation. Changes in gene expression are often caused by mutations in *cis*-regulatory elements (CREs) and trans-regulatory elements (TREs) (7). *Cis* and *trans*-regulators control transcription of genes located on the same chromosome, or on both homologous chromosomes, respectively. Several recent technological breakthroughs (7–10) have considerably improved our ability to study these regulatory sequences and their associated expression profiles. They revealed that gene expression profiles were highly variable and heritable within species (2,11,12), quickly diverging among species (8). Furthermore, many studies have shown how changes in gene expression contribute to adaptive changes, by natural selection at the individual level (4,6,13–17). From a theoretical standpoint, several models have also been developed to understand the evolution of gene expression by individual level selection (18–22).

In this paper we investigate a new and different selective phenomenon also acting on the evolution of regions controlling transcription initiation. This phenomenon may contribute to the fast divergence of regulatory networks between closely-related species, but unlike the usual view, it is not rooted in individual-level selection, and hence does not necessarily increase individual fitness. We term this selective process ‘ER‘ (for ‘enhancer runaway’). It results from competition for expression between homologous cis-acting regulatory sequences in diploid organisms. With this gene-level selection, these regulatory sequences behave as self-promoting genetic elements.

Transcription initiation is determined by the binding of a suitable RNA-polymerase to a Transcription Start Site (TSS). This binding involves a complex machinery of often over 30

partner proteins (23). It depends on the interaction of CREs and TREs. CREs are non-coding sequences located on the same chromosome as the regulated gene. They include core promoters located around the TSS, which integrates the regulatory inputs (24). They also include enhancers, which influence transcription initiation rate independently from their orientation or localization on this chromosome (9). TREs are coding sequences, located anywhere in the genome, that produce transcription factors (TFs), which bind to CREs on both homologs. They can also produce cofactors, which bind other proteins (including other TREs) (7).

In order to show how the ER process works, we use population genetic models with both protein-coding sequences (the 'gene(s)') and regulatory sequences (enhancers or transcription factors). For generality, we do not specify precisely how regulatory mutations change regulatory networks. We simply assume that mutations occur in enhancers and TFs, which impact expression levels. Indeed, there are many ways for an enhancer mutation to modify expression levels (2,7,9,25,26). It may change for instance the binding site affinity towards TFs, the number and/or spacing between binding sites, or the nucleosome conformation around the enhancer region, which strongly impacts DNA accessibility for the transcriptional machinery. Because enhancers act in *cis*, the presence of different alleles at the enhancer locus creates a pattern of allele-specific expression (imbalance in chromosomal origin of transcripts) for the protein-coding gene (27,28). Indeed, in heterozygous individuals, the 'weaker' enhancer contributes less to protein expression than its 'stronger' counterpart on the other chromosome, causing imbalanced expression of homologous gene alleles. TF mutations may also alter expression levels in various ways. For instance, they can change its DNA-binding domain (modifying its affinity for different binding sites) or its protein-protein binding domain (modifying its interactions with e.g. cofactors or histones). However, they act in *trans*, and do not generate allele-specific expression.

To obtain a good understanding and broad evaluation of the significance of the ER process, we investigate several complementary models. In a first model, we study the ER process in the absence of individual level selection for expression level. We then incorporate individual level selection on expression. Much debate has been going on over the different selection pressures acting on expression levels. While some argue that regulatory polymorphism is mainly neutral or quasi-neutral (29,30), others suggest that regulatory evolution is mostly

shaped by stabilizing selection (31–33) and occasionally by directional selection (8,34–36). In a second and third model, we introduce stabilizing selection on expression levels (due to a relative gene dosage constraint in model 2, or an absolute expression constraint in model 3). In these two models, other regulatory regions can evolve in concert (other enhancers in model 2, TFs in model 3). Finally, since the ER process may act very differently in genomes exhibiting inbreeding and low heterozygosity, we investigate how it is influenced by the mode of reproduction (inbreeding in model 4). Overall we show that, in a small genomic region around genes, there is a selection pressure on enhancer to increase expression levels. This phenomenon leads to an open-ended escalation in enhancer strength (a 'runaway'). This process is not halted by inbreeding, or by stabilizing selection on expression levels, as long as enhancers can evolve in concert with other regulatory sequences (enhancers or TFs) involved in the same regulatory network. Enhancer runaway is not a highly specific or idiosyncratic process: it is expected to occur at variable intensities for all genes in nearly all eukaryotic diploid organisms. This widespread phenomenon may significantly shift our current understanding of gene regulatory regions, and opens a wide array of possible tests and comparisons.

Model

Competition for expression

To illustrate how competition for expression works, we first present a two-locus model in an infinite, diploid, sexual population. The first locus, the 'gene', codes for a protein. We suppose that it undergoes recurrent deleterious mutations (with fitness effect s and dominance h) at a rate u, but the argument would apply equally well with beneficial mutations (see methods). This locus quickly reaches the usual deterministic mutation–selection equilibrium, with deleterious mutation at frequency u/hs. The second locus, the 'enhancer', is located at a recombination distance r and controls the expression of the gene in a cis-regulatory fashion. We wish to determine the selection pressure acting on mutations that modify the strength of this enhancer.

In model 1, we consider that the overall expression level is tightly controlled, for instance because of trans-acting regulatory factors producing a negative feedback loop. Such a feedback loop is not relevant to all genes, but is a particularly useful starting case. It allows investigating how the ER process works without the complication of additional selection pressures acting, at the individual level, on protein expression. We thus assume that overall expression levels are constant (due to the feedback loop), such that only the relative contributions of each homologous alleles vary due to mutations on enhancers.

Labelling e_1 and e_2 the strengths of the two enhancer alleles of an individual, the gene associated with enhancer of strength e_1 contributes a fraction $e_1/(e_1+e_2)$ of proteins produced. As a consequence, the gene allele linked to a stronger enhancer contributes a larger share of proteins.

In double heterozygotes, if the deleterious allele of the gene is linked to the weaker enhancer, less than 50% of the proteins produced will be of the defective form, and thus its effect on fitness will be less than predicted based on its dominance coefficient h (which in effect reduces its dominance to $h_1<h$). In contrast, if it is linked to the stronger enhancer, its deleterious effect on fitness will be stronger than predicted by h (which in effect increases its dominance to $h_2>h$). The fitness effect of linkage with a specific enhancer can thus be

modeled as a change in *h* that occurs only when the enhancer is heterozygote, while dominance is necessarily *h* as long as an enhancer allele is fixed, as illustrated in Fig 1.

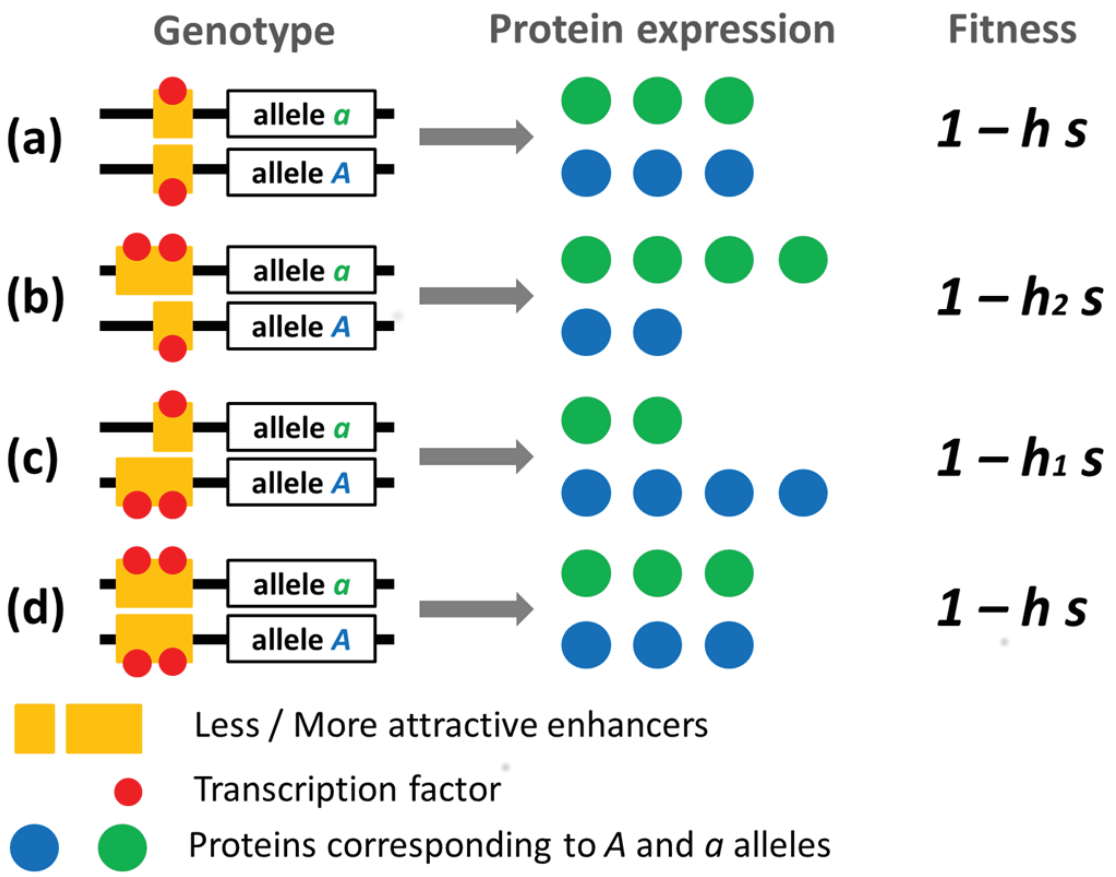


Figure 1 - Schematic representation of protein expression and fitness of gene heterozygotes. Here, the strengths of two enhancer alleles are represented by their ability to attract transcription factors. Four genotypes are represented: weaker enhancer homozygote (a), stronger enhancer homozygote (d) and enhancer locus heterozygotes (b) and (c). In enhancer locus heterozygotes, the stronger enhancer is either associated with the deleterious gene allele (b) or with the viable gene allele (c). Corresponding fitnesses are indicated. Note that we consider here a case where the total amount of proteins produced is constant.

Because deleterious mutations are usually partially recessive (37), the relationship between the fraction of defective proteins and fitness is necessarily nonlinear, monotonously decreasing and concave (Fig 2). Thus, we have $h_1 + h_2 > 2h$ (applying Jensen's Inequality to the strictly concave fitness function). In other words, a polymorphism at the enhancer locus necessarily increases average dominance (the mean of h_1 and h_2 is greater than h), which increases the strength of selection against heterozygous deleterious mutations, but also reduces the fitness of individuals carrying these mutations.

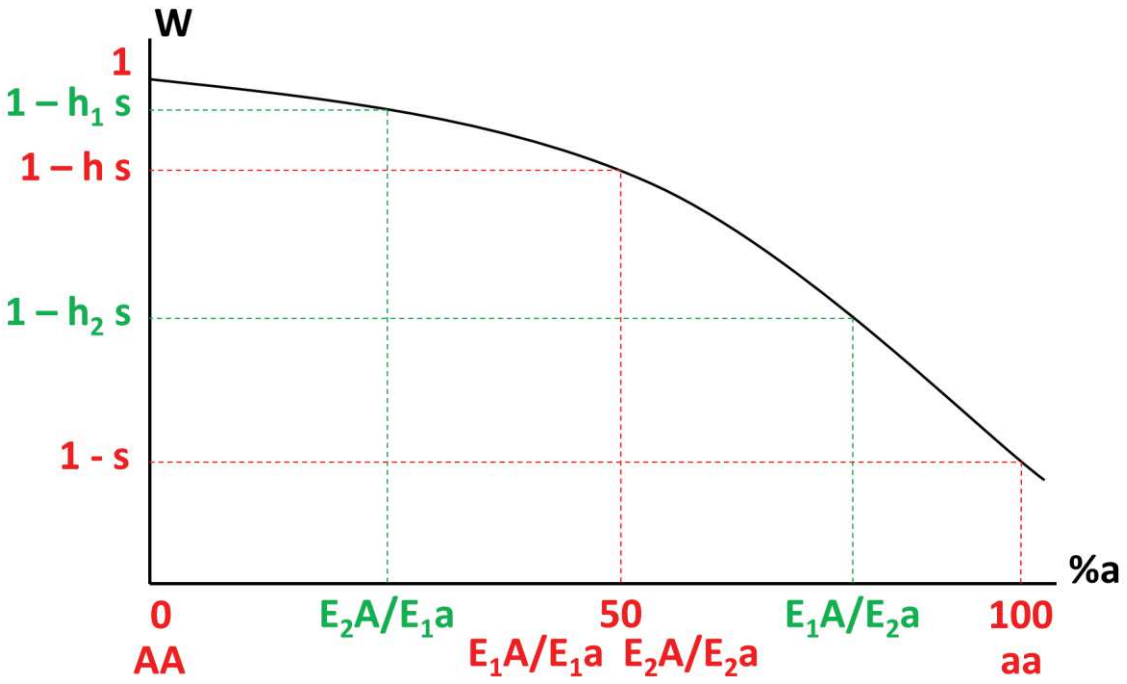


Figure 2 – Variation of fitness (W, y-axis) as a function of the percentage of defective proteins noted %a (x-axis, from 0% in AA homozygotes to 100% in aa homozygotes) in different genotypes, where E1 and E2 are the weaker and stronger enhancer alleles respectively ($\mathbf{e_2} > \mathbf{e_1}$). This function is necessarily monotonic and concave when deleterious mutations are recessive (h < ½).

The deterministic change in frequency at the enhancer locus can be computed from standard population genetics equations, considering a generic diploid life cycle with four steps: diploid selection, meiosis with recombination, mutation and syngamy. This frequency change can be decomposed into two terms that correspond to 'masking' and 'purging', as introduced in previous related models (38,39) (see derivation in methods). The 'masking' term does not depend on recombination, and is frequency-dependent. It always disfavors the enhancer allele when rare, independently of its strength. The reason is that a rare enhancer will most often be represented in double heterozygotes (since deleterious mutations often are rare too), which have lower average fitness, since they have higher average dominance (this is similar to Fisher's argument for the evolution of lower dominance (40)). In other words, rare enhancer alleles pay the cost of unmasking deleterious alleles. This term is strongly conservative, as it prevents any new enhancer to enter the population. The 'purging' term, in contrast, is frequency-independent, always favors stronger enhancers, and increases when recombination between the enhancer locus and the gene decreases. The reason is that genes that are linked with stronger enhancers are more exposed to selection and more efficiently purged from deleterious mutations. Hence, stronger enhancers are disproportionately found on –and hitchhike with– favorable genetic backgrounds. Overall, combining these two effects, weaker mutant enhancers are always disfavored, while stronger mutant enhancers are favored if sufficiently tightly linked to the gene for the 'purging' effect to overcome the 'masking' effect. The recombination distance where the two effects balance each other depends however on the strength difference of the two enhancer alleles. For enhancers that are very similar in strength, the range of recombination distances where the runaway can occur is larger, but the intensity of selection on the stronger enhancer is weaker (see Fig S1). This is due to the fact that, for small differences in enhancer strength, Δh becomes vanishingly small compared to δh (and thus the masking term in front of the purging term, see eq. (3) and (5) in methods). Overall, stronger runaway is expected close to genes because selection intensity on new enhancers is stronger at small recombination distances and because the runaway only concerns enhancers of small effects at larger recombination distance.

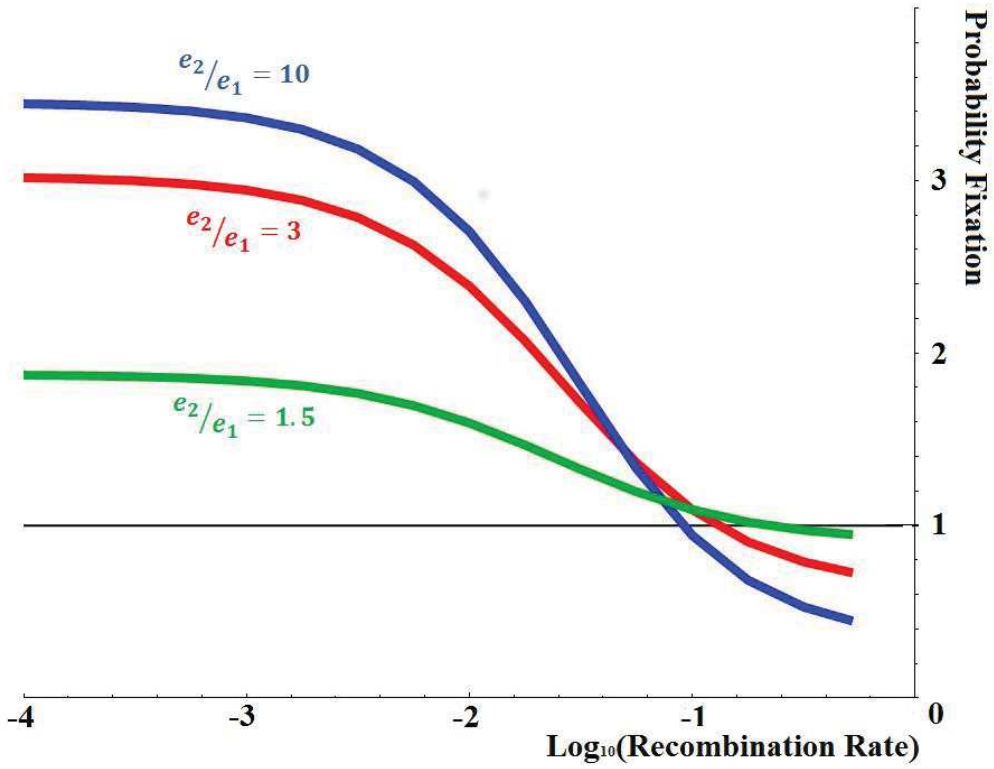


Figure S1 – In blue, the mutant enhancer (e_2) is ten times stronger than the wild-type enhancer (e_1). In red, it is three times stronger. In green, it is 1.5 times stronger. This figure was obtained using the analytical version of model 1 (see Methods), with following parameters: partial recessivity $h = 0.25$, selection intensity $s = 0.1$, N_{pop} $u = 1$. Results show that, at short recombination distances, selection for stronger enhancers is larger for larger enhancer strength differences. However, at larger recombination distances, selection for stronger enhancers decreas faster for larger strength differences. Consequently, the recombination rate limit after which stronger enhancers are selected against is larger for smaller enhancer strength differences.

In regulatory regions close to the gene, selection thus favors enhancers contributing more to protein production: homolog enhancers compete for expression. This outcome is illustrated on Fig 3, where this analysis is checked against –and agrees with– stochastic numerical simulations reporting the fixation probabilities of new enhancers in finite populations.

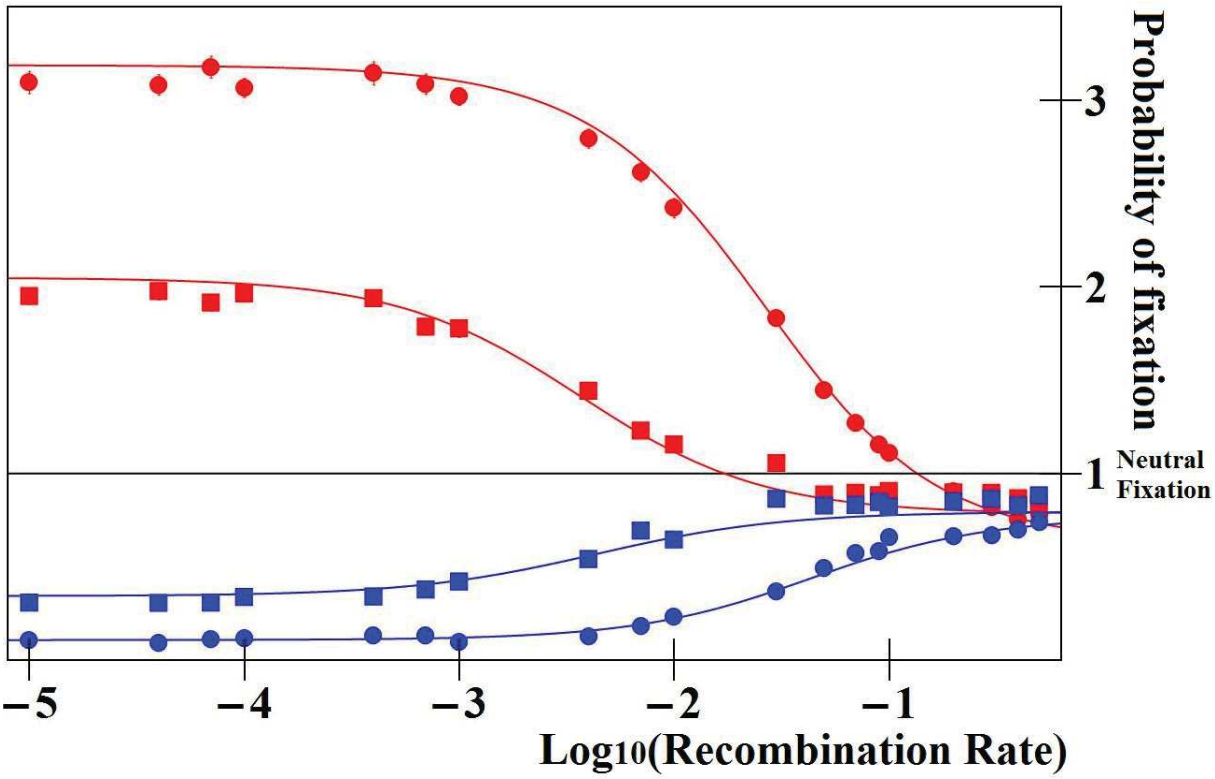


Figure 3 – Ratio of fixation probabilities of mutations altering enhancer strength relative to that of a neutral mutation. In red, the mutant enhancer is three times stronger than the wild type; and in blue, three times weaker. Simulated and analytically predicted values (see methods) are represented by dots and lines, respectively. Values are reported for the case where enhancer strength evolution does not alter overall protein expression (model 1, see text), for various recombination rates between the enhancer and the gene (x-axis), for weak (s = 0.01, squares) or strong selection (s = 0.1, circles) and partial recessivity (h = 0.25). Results are illustrated for N_{pop} u = 1 (where N_{pop} is the population size and u the gene mutation rate). Results for higher N_{pop} u just need to be multiplied by a factor equal to N_{pop} u. For instance the relative fixation probability for N_{pop} u = 10 of tightly linked stronger enhancers would be ~20 or ~30, for s = 0.01 or s = 0.1, respectively. Weaker enhancers (blue) are always selected against, while stronger enhancers (red) are selectively favored provided that they are closely linked to the gene, and disfavored otherwise.

This finding indicates that stronger and stronger cis-regulatory elements should evolve in an open-ended fashion in the vicinity of genes, as long as that does not influence the total amount of proteins produced. Competition for expression may thus be responsible for a runaway process of enhancer strength. During this process, allele-specific expression would transiently occur, but expression balance would be restored once the new enhancer reaches fixation. This process shares many similarities with the endless occurrence and spread of

new segregation distorters that transiently bias Mendelian ratio while they sweep (41). It also shares similarities with models of Fisher runaway in the context of sexual selection, where female mating preference drives the open-ended evolution of extravagant traits in males (42).

Interestingly, the spread of such stronger enhancers occurs even though it temporarily decreases population mean fitness (see methods). Indeed, stronger enhancers spread because they find themselves on better backgrounds, but at the expense of temporarily increasing mean dominance and hence unmasking deleterious mutations. Overall the ER process does not optimize mean individual fitness in the population.

Competition for expression and stabilizing selection on expression levels

Gene expression regulation is not necessarily embedded in a negative regulatory loop, as considered above. Mutations on enhancers will often alter overall expression levels. When this occurs, stabilizing selection on overall gene expression will interfere with competition for expression. We developed two additional models (model 2 and 3) with such interactions (see methods). In these models, mutations on enhancers alter both relative contribution to expression, $e_1/(e_1+e_2)$, and total expression levels, e_1+e_2. We assume that total expression levels undergo stabilizing selection with different intensities that reflect the functional diversity of genes (over- or under-expression may be more costly for some genes than others).

In model 2, we assume that stabilizing selection on expression levels stems from gene dosage, such that the optimal amount of a given protein depends on the amounts of another protein coming from another loci, as occurs for instance in enzymatic or metabolic pathways. This produces the strongest constraint on enhancer strength evolution when only two proteins are concerned (see methods). In this case, any increase in the expression of gene 1 causes a departure from optimal dosage for gene 2, which reduces fitness. Results show that such stabilizing selection fails to prevent enhancer strengths from escalating. This is because enhancers of both genes coevolve: their strengths increase in parallel, allowing maintenance of the correct protein dosage. However, stabilizing selection tends to decrease escalation rates (longer doubling times on Fig 4, see methods). Strength-increasing mutations of large effects are indeed counter-selected, since they lead to large deleterious

departures from optimal gene dosage. However, strength-increasing mutations of small effects lead only to small enough departures from optimal gene dosage for the genotype to survive until optimal gene dosage is restored by compensatory mutations. Stabilizing selection needs to be relatively strong (stronger than the selection at the gene locus) for it to significantly alter escalation rates.

In model 3, we considered a situation where stabilizing selection acts directly on the absolute expression levels of a gene, but transcription factors influence expression levels as well as enhancers. We designed a three-locus model with a gene, an enhancer locus, and a TF locus, where both the strength of the enhancer and TF combine to determine expression levels (see methods). Here again, as shown on Fig 4, stabilizing selection slows down but does not stop the ER process, due to coevolution between regulators (here, enhancers and TFs). As enhancer strength increases, TF strength decreases in proportion, which maintains approximately constant and optimal total expression levels. In this model, escalation rates are systematically lower than the ones obtained with model 2 (except for $I = 0$). This is to be expected as two identical enhancer loci are exposed to the ER process in model 2, while only one enhancer is exposed to the ER process in model 3 (the TF locus only responds to stabilizing selection). As a consequence, in model 3, stabilizing selection importantly decrease escalation rates at lower intensities (lower than the intensity of selection at the gene locus).

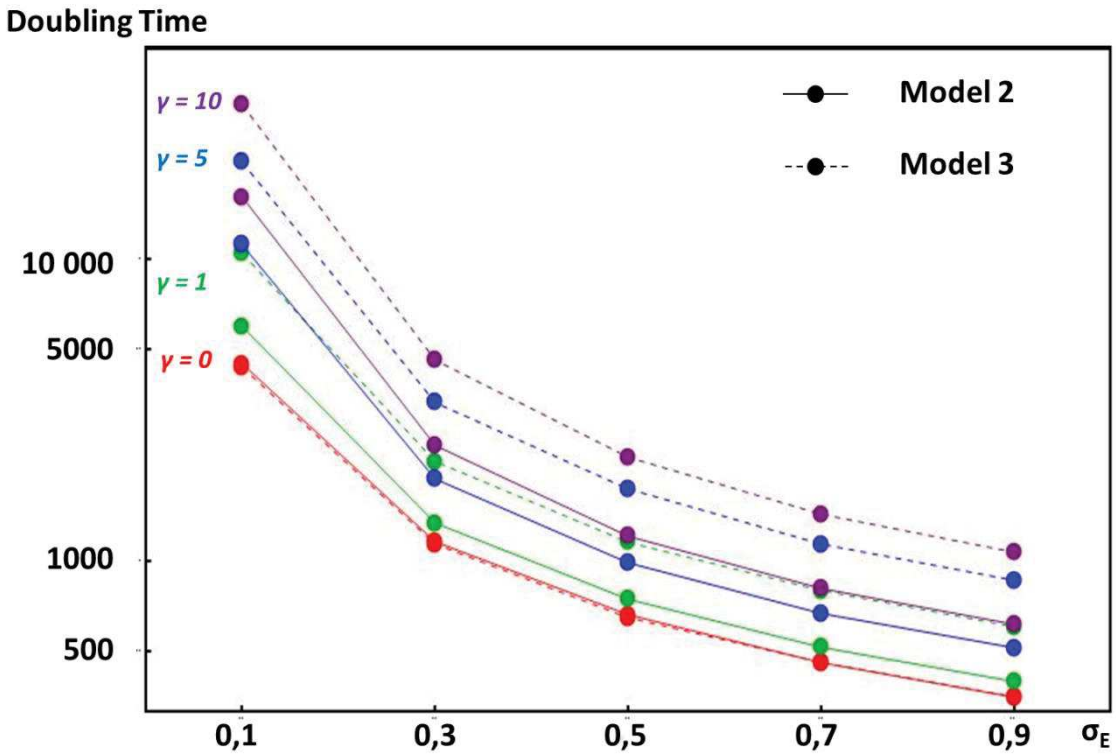


Figure 4 – Comparative doubling times of enhancer strength escalation in models considering different selection pressures acting on overall gene expression. Y-axis indicates doubling times of enhancer strength (the expected number of generations needed to double the initial enhancer strength) according to the mutation size standard deviation on enhancers (x-axis). Because in all models presented, enhancer strength increases open-endedly on average, this doubling time measures the rate of escalation (see methods). The mutation size standard deviation on enhancer strength (σ_E) is a measure of the magnitude of mutational input (see methods). In model 2 (plain lines), overall expression of one gene is involved in a dosage relationship with the overall expression of another gene. Overall expression is determined by one enhancer locus per gene, and any departure from optimal dosage is costly. In model 3 (dashed lines), the absolute amount of protein produced is under stabilizing selection, but expression level is influenced by both an enhancer and a TF locus. In both models, stabilizing selection intensity is scaled by the intensity of selection at the gene locus (see methods). Deceleration of ER process due to stabilizing selection is illustrated for γ values equal to 0 (in red), 1 (in green), 5 (in blue) and 10 (in purple). Doubling times were obtained using stochastic simulations (see methods).

Impact of mating system on competition for expression

Variation in enhancer strength only makes a phenotypic difference in double heterozygotes. In situations where such heterozygotes are less frequent than under random mating, the pace of the ER process is expected to be slower. Inbreeding is a typical and common situation decreasing heterozygote frequency. It may be caused by various processes (e.g. selfing in hermaphroditic plants or animals, population structure). To determine whether ER indeed differs among species exhibiting e.g. different mating systems, and to quantify the extent of this reduction, we introduced partial selfing in the first model.

In model 4, at each generation, individuals were considered to have probability p_s of selfing (i.e. with probability p_s, the same parent is used to sample the second gamete). As expected the ER process slowed down as selfing rate increased. Fig 5 illustrates this behavior. Results show that relatively high levels of self-fertilization are needed for the ER process to be significantly slowed down.

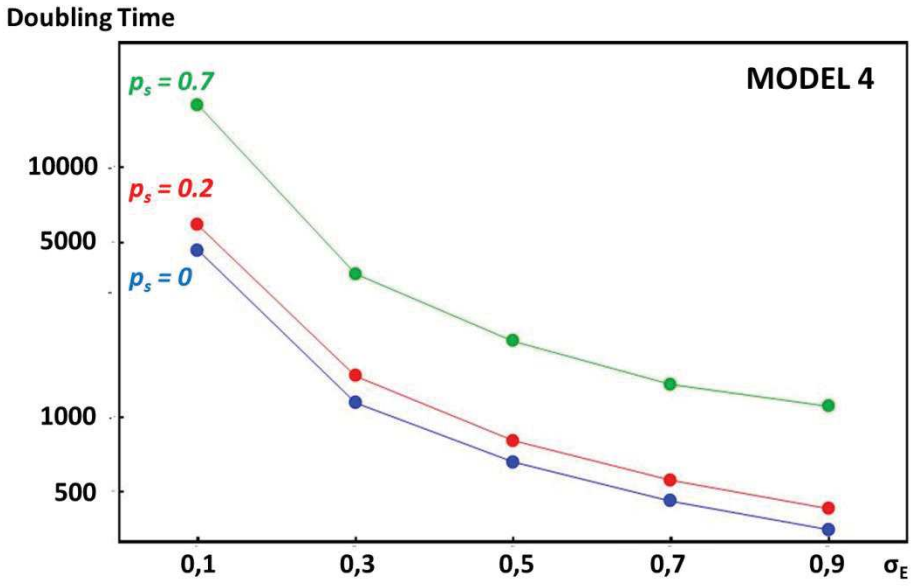


Figure 5 – Comparative rate of enhancer strength escalation in models with or without self-fertilization. Axes and simulation methods are the same than for Fig 4. Selfing rate p_s is 0, 0.2 and 0.7 on blue, red and green curves, respectively. In all three cases, enhancer strength escalation is faster with larger mutational variance σ_E^2 on the enhancer locus. Self-fertilization slows down the ER process.

Discussion

The ER process we describe sheds a new light on the evolution of gene regulatory regions in diploids. It should be a widespread phenomenon, applicable to all genes in diploid eukaryotes as the theory only involves generic assumptions (recessivity of deleterious mutations (37), genetic variation at enhancer loci (43), CREs - TREs coevolution (44), stabilizing selection on expression levels (31–33)). This process has been acting probably since early unicellular diploids, more than a billion years ago, before the evolution of complex combinatorial or developmental regulatory pathways. As we already stressed in the introduction, it does not undermine the idea that expression regulation is also subject in contemporary eukaryotes to other important selective effects (direct selection for expression level, timing and localization).

Before discussing the implications and predictions associated with this theory, it is useful to relate it to previous models involving variation in dominance, ploidy, mutation load and selection for modifiers. It shares with models of ploidy evolution the fact that increased haploid expression leads to more efficient purging of deleterious mutations, which benefits tightly linked modifiers (38,39). It shares with models of dominance evolution (45–48) the limit that selection on modifiers is weak, of the order of the mutation rate (even if this issue may be alleviated if migration is the source of deleterious alleles (48,49)). Such a weak selection is usually seen as a limit in the case of dominance evolution because genetic drift in small populations or small pleiotropic effects of modifiers should easily overwhelm the selection pressure for dominance modification (50). A distinctive feature of our theory is that the ER process is not strongly limited by such pleiotropic effects: as we showed, direct selection against suboptimal expression levels can be readily compensated by CREs-CREs or CREs-TREs coevolution, without halting the ER process. A critical difference with previous models is that dominance does not evolve: the dominance of deleterious mutations is h before the sweep of a stronger enhancer, and is still h after its fixation. Similarly, ploidy or the number of gene copies stays constant and does not evolve (with the consequence that, unlike in models involving gene duplicates or ploidy variation, the mutational load is not permanently changed by a change in gene number (38,47)). In our models, selection is frequency-dependent at leading order and leads to a runaway escalation not seen in models

of ploidy or dominance evolution. Our models specifically focus on cis- or trans-acting modifiers, mimicking the actual genetic variation occurring on regulatory regions. Cis-regulation introduces naturally an asymmetry in the fitness of the two double heterozygotes (EA/ea and Ea/eA), which is also not a feature present in models of ploidy or dominance evolution.

This theory leads to a series of predictions that can be further tested. We highlight eight of them below using capital letters A-H. The ER process should occur for almost all genes in diploids, but with different, and possibly very different rates depending on their specific evolutionary constraints (regulatory loops, dosage relationships, intensity of stabilizing selection). For instance, as we illustrated with our first model, this runaway is fastest for enhancers that are embedded in a downstream negative regulatory loop (prediction A). Such negative feedback loops have been extensively described and are often thought to largely contribute to phenotypic robustness (e.g. (51–53)). This runaway should be slower for genes that are not regulated with such loops and exposed to direct stabilizing selection on relative or absolute expression levels. The pace of the ER process depends in those cases on the form and intensity of the stabilizing selection, as well as on the opportunities for CREs and TREs to coevolve to maintain optimal total expression levels, which may differ among genes.

Our results show a strong impact of the recombination rate between the enhancer and the gene on the ER process (see Fig 3). There are two regimes: (1) for large recombination rates, the ER process does not occur (large and medium effect enhancers are selected against and small effect enhancers are nearly neutral), while (2) at shorter genetic distances, the runaway occurs and its rate increases as the recombination rate decreases. Everything else being equal, the ER process should cause a positive correlation between enhancers' strength and their proximity to the gene, but it should not concern enhancers located at large genetic distances (prediction B). This prediction would be consistent with the observation that most CREs remain close to the gene they regulate (54). The critical genetic distance delimiting the two ER regimes increases as the intensity of selection on the gene increases. As a result, genes undergoing stronger purifying selection would be expected to have a larger surrounding region where enhancers increasing expression may arise and compete for expression. Consequently, we predict that genes experiencing stronger purifying selection

should exhibit a larger surrounding regulatory region, and that, for a given genetic distance to such genes, enhancers should exhibit a faster ER process (prediction C). Such qualitative predictions may be altered if there is an inherent physical tendency for CREs strength to covary with physical map distance. However, if, as is most likely, CRE strength decreases with physical distance along the chromosome, the qualitative pattern will remain identical (since decreased CRE strength and increased recombination both prevent the fixation of new enhancers).

There are several reasons, not included in our models, why the ER process could slow down or stop. First, many mutations on enhancers are likely to have pleiotropic effects on timing or localization of expression. These features are not included in our model, and probably limit the availability of enhancer mutations that can contribute to the ER process in higher eukaryotes. Second, mutations may not be able to increase enhancer strength indefinitely. For instance, when a binding motif sequence is optimized for a particular TF, there may be little room for further improvement. Similarly, there is a limit to the number of binding motifs that can be packed in a regulatory region, etc. These constraints could be revealed by studying the effect of random mutations on regulatory regions. If enhancer strength is maximized by the ER process, no (or very few) random mutations on the enhancer will be able to increase its strength, such that the strength of those enhancers should be biased downward by random mutations. Conversely, if a TF has evolved to compensate for the enhancer runaway (i.e. by increasingly repressing transcription), random mutations on this TF should on average increase expression levels (prediction D).

The evolution of regulatory networks' complexity is not well understood and controversial (some argue that the evolution of complex regulatory networks stem from adaptive processes (18,20,55), while others consider a non-adaptive origin (56,57)) . We saw that the ER process depends on the shape of regulatory networks (e.g. presence of feedback loops, of gene dosage etc.). It may also be facilitated with more complex regulatory architecture, as highly degenerate and complex regulatory architectures (58) could provide more 'degrees of freedom' for CREs - TREs (co)evolution. Interestingly, the ER process may also contribute to the evolution of complex regulatory networks. For instance, an increase of enhancer strength may be achieved e.g. by locally duplicating TF binding motifs (and hence increasing

the complexity of regulatory architecture). The duplicated motifs may further diverge to attract a larger diversity of TFs if this happens to be a route to further increase of enhancers' strength: combinatorial regulation (where expression specifity results from a particular combination of several TFs, as in e. g. (59)) may thus evolve from the ER process. A positive feedback loop may thus occur between the ER process and the evolution of regulatory complexity. Much of these effects will rely on the contingency of mutational variation on enhancers, but can produce a level of architecture complexity much beyond the level expected under individual selection alone. Species where the ER process is expected to be faster (e.g. outcrossing vs. selfing diploids, diploid versus haploid eukaryotes) should exhibit more complex regulatory architectures (prediction E). Like for other theories for the evolution of complexity (56,57), complexity would emerge here as a by-product of another process (here ER) and is not a direct target of selection.

It is generally envisioned that many regulatory networks can be functionally equivalent. For instance, many CREs - TREs combinations can perform the same signaling, and different global regulatory wirings among TREs and CREs can achieve the same regulatory pattern. Like with a key / lock or signal / receiver mechanisms, the central functional requirement is the reciprocal recognition of interacting regulatory elements. Such a situation produces a fitness landscape with a ridge, along which 'evolutionary freedom' allows for substantial neutral divergence (60). This process is often thought to drive relatively fast divergence of regulatory networks among species without altering much expression patterns (61). In our models, the same process occurs (coevolution between regulators occurs without modifying expression levels much), but, in addition, selection for stronger proximal CREs leads to faster-than neutral divergence (prediction F). Using again the fitness landscape metaphor, it means that the ridge is actually not flat, but is slightly sloping: evolution between networks with the same expression pattern is not neutral, but rather directed to favor networks involving stronger proximal enhancers. Moreover, this faster-than-neutral divergence is expected to be accentuated for CREs (compared to TREs, prediction G). Indeed, the ER process necessarily involves CREs, and optionally TREs. For instance, cis- cis- coevolution illustrated in model 2 does not involve TREs evolution. This may partly explain why CREs are usually (but not always, e.g. (62)) found to contribute more than TREs to expression

regulation divergence among species (e.g. in *Saccharomyces* (63), *Drosophila* (64), *Gasterosteus* (16) and *Mus* (65)).

As shown in model 4, the ER process is slower in presence of inbreeding. In particular, the ER process is expected to be faster in outcrossing than in self-fertilizing species (prediction H). One way to test for this prediction would be to compare CRE strengths in hybrids between closely related outcrossing and self-fertilizing species. Indeed, in the hybrid, expression level differences between alleles can only be explained by CRE variation, as TREs from both parents are shared (66). Such method could be used in e.g. Arabidopsis (67) or Capsella (68). The expectation would be that CREs derived from the self-fertilizing species should be weaker on average, biasing expression pattern in F1s towards the outcrossing parent alleles. Such test would require however to control for the direction of the cross (distinguishing the species effect from maternal versus paternal effects).

Several genetic oddities that have been observed in some species may greatly limit the ER process. For example, in somatic cells of Diptera, homolog chromosomes pair and expression is largely influenced by a phenomenon referred to as 'transvection' (69). With transvection, "CREs" impact regulation of both homolog chromosomes (i.e. are not really behaving as cis-regulatory elements as they also regulate in *trans*). Somatic pairing and transvection in these taxa certainly reduce, or even eliminate competition for expression. Similarly, repulsion of homologs, as found in mammals' nucleus (70), (i.e. the fact that the chromosome territories of homologs tend to be further apart within the nucleus compared to a random pair of chromosomes) could also strongly reduce competition for transcription factors between competing CREs, as the transcription of homologs likely involves different and spatially segregated 'transcription factories' (71). In both cases, reduced competition for expression between homologs could strongly limit the ER process. Finally some recombination hotspots are located close to genes in several species (72,73). This could also strongly limit the ER process by breaking linkage disequilibria between CREs and genes. Whether some of these genetic oddities evolved as suppressors of the ER process is an intriguing possibility, especially given that their evolutionary significance remains elusive.

Methods

We describe below the models (1-4) we designed to infer enhancer strength evolution. These models differ by the mode of reproduction and the type of stabilizing selection acting throughout gene regulatory networks. In model 1, we study the evolution of enhancers assuming that expression levels are so tightly controlled that the total amount of proteins is strictly constant. There is no selective pressure resulting from changes in overall protein expression. In such a situation, a competition for expression appears between homolog enhancers. We used this model to understand the genetic cause of the resulting ER process and the evolutionary patterns it creates. While this model is relevant for genes with regulatory loops (and for understanding selection pressure acting on expression balance), it may not capture all selection pressures acting on enhancers. Thus, in following models (2 and 3), we consider that enhancers do influence the total amount of protein expressed and investigate how it interacts with the selection pressure described in model 1. We consider two cases. In model 2, there is an optimal dosage between expression levels of different genes: enhancer strength variation on one gene causes a departure on gene expression dosage, which is deleterious. In model 3, the fitness penalty arises from any departure from an absolute optimal expression level, but expression levels depend on enhancers and transcription factors, and not only on enhancers as in previous models. Finally in model 4, we introduce partial selfing in a model 1 genetic setup, to infer the effects of inbreeding on the ER process.

Designing model 1

We firstly consider a diploid two-locus model: the first locus, referred to as 'the gene' is a protein-coding locus undergoing mutations and diploid viability selection; the second one, referred to as 'the enhancer', is a CRE locus controlling the expression of the gene. The gene is exposed to recurrent deleterious mutations, at a rate u per individual per generation, changing A alleles into a alleles. In the analytical derivation below, we focus on this bi-allelic case, but this assumption is relaxed later for numerical simulations. We define the relative fitness of genotypes as *1*, *1 – h s* and *1 – s* for *AA*, *Aa* and *aa* genotypes, respectively, where *s* is the selection coefficient against the *a* allele and *h* its dominance. The enhancer is at a recombination distance *r* from the gene. We consider two alleles (E_1 / E_2), which differ by

their ability to promote expression of the gene located in *cis* (i.e. on the same chromosome). This ability is referred to as their 'strength' and noted e_1 and e_2. Biologically this strength depends on many parameters, such as sequence affinity, chromatin state, binding network or intracellular signals (7). Furthermore, different mechanistic models have been put forward to describe enhancer effects: they are thought to change either promoter activation levels or the probability that a promoter will be activated to initiate transcription (74). Here, we will not assume a particular mechanism but cover all these possibilities by simply supposing that enhancer sequences intrinsically differ by their ability to activate gene transcription.

We assume that the gene on the same chromosome than an enhancer with strength e_1, facing on the homolog chromosome an enhancer with strength e_2, contributes to a fraction $e_1/(e_1 + e_2)$ of protein produced (see Fig 1). As a consequence, the gene associated with a stronger enhancer contributes a larger share of proteins. A major feature of this first model is that the total protein expression level is tightly regulated: the total amount of protein produced is assumed to be constant, and does not depend on enhancers' strength (6 proteins are always produced on Fig 1). Such a situation would occur, for instance, when the amount of transcription factors, or a repressor is downstream regulated with a negative feedback loop (e.g. through the total amount of protein produced or by the concentration of a downstream metabolite resulting from protein activity). Besides representing this fairly common biological situation, this model is also important to understand the evolution of enhancer strength independently from the selection pressure acting on the amount of protein in a given cell at a given time.

When the gene locus is homozygous, the fitness of individuals does not depend on the enhancer alleles, as 100% of the proteins are of the same type (the fitness is *1* or *1 – s* for *AA* and *aa* genotypes respectively). However, different situations arise in individuals that are heterozygous at the gene locus. When they are homozygous at the enhancer locus, they equally express each type of proteins (50% each), and their fitness is *1 – h s* (genotypes (a) and (d) in Fig 1). When they are heterozygous at the enhancer locus they express more (or less) of the defective protein if the stronger (or weaker) enhancer is associated with the deleterious allele (genotypes (b) and (c) in Fig 1). This changes dominance at the gene locus. When fewer defective proteins are produced, the deleterious effect will be lower, which

means lower dominance, noted h_1 such that $h_1 < h$ (genotype (c) in Fig 1). Conversely, if more defective proteins are produced, the deleterious effect will be larger, which means a higher dominance, noted h_2 such that $h_2 > h$ (genotype (b) in Fig 1). The values of h_1 and h_2 will depend on the strengths e_1 and e_2 of E_1 and E_2 alleles. In order to be more specific about this relationship, we note that phenotype-to-fitness relationship must verify two major properties: (1) the fitness must decrease as the proportion of deleterious proteins expressed increases and (2) the relationship between the fitness and the proportion of defective proteins expressed must be concave, as deleterious mutations are most often partially recessive (37), which means that fitness effects of deleterious mutations are lower than what would be expected in an additive (linear) situation. As a consequence, the relationship between fitness and the fraction of defective proteins expressed must be similar to the situation illustrated on Fig 2. A generic way to formulate these properties is to express fitness as a function of $f_{[a]}$, the fraction of defective proteins, h and s with a concave monotonic power function:

$$W = 1 - s f_{[a]}^{-\frac{Log(h)}{Log(2)}}, \tag{1}$$

which conveniently converges to the additive situation for $h = 1/2$. With such a relationship, and assuming that $e_2 > e_1$, we have:

$$\begin{cases} h_1 = \left(\dfrac{e_1}{e_1 + e_2}\right)^{-\frac{Log(h)}{Log(2)}} & \text{(2.a)} \\ h_2 = \left(\dfrac{e_2}{e_1 + e_2}\right)^{-\frac{Log(h)}{Log(2)}} & \text{(2.b)} \end{cases}$$

Noting $\Delta h = (h_1+h_2)/2 - h$, one can use Jensen's Inequality on the strictly concave fitness function to show that $\Delta h > 0$. The commonly accepted assumption that deleterious mutations are partially recessive (i.e. h is on average around 0.25 for mildly deleterious

mutations, (37)) leads in our model to the outcome that polymorphism at the enhancer locus increases the mean dominance.

Derivation of mutant enhancer frequency change over one generation

To study the evolution of enhancer strength, we are first looking for the frequency variation of the stronger E_2 enhancer allele after one generation (noted as Δp). To do so, a four-step life cycle is implemented, with diploid selection, meiosis, mutation and syngamy. The exact recursion is then linearized to leading orders, defining a parameter $\xi \ll 1$ and assuming (1) weak selection on the gene (*h s* is of order *ξ*), (2) very small mutation rate (*u* is of order ξ^2) and (3) small recombination rate *r* between enhancer and gene loci (of order *ξ*). Under those fairly generic assumptions, the frequency of the deleterious allele *a* (noted p_a) is small (of order *ξ*) and the linkage disequilibrium (D_{EA}) between the enhancer and the gene is small (of order *ξ*). D_{EA} is defined as positive when E_2A and E_1a haplotypes are over-represented (meaning that D_{EA} is positive when stronger enhancers are associated with beneficial alleles). We obtain, noting *p* the frequency of E_2 and *q* the frequency of E_1, the leading order of the frequency variation of E_2:

$$\Delta p = -2\Delta h p_a pq(1-2p)s \tag{3.a}$$

$$+ D_{EA}[4hpq + (1-2p)(h_2 q - h_1 p)]\, s + o(\xi^2) \tag{3.b}$$

The first term (3.a) corresponds to a direct selection on enhancers while the second term (3.b) is an indirect selection proportional to D_{EA}.

The direct selection is negative when E_2 is rare (since $\Delta_h > 0$) and positive when it is frequent. Thus, it favors the most common allele at the enhancer locus. To understand this effect, it is useful to derive the mutation-selection equilibrium of p_a, p_a^{Eq}. Using the same linearizing assumptions than for Δp, assuming that the mutant stronger enhancer has just entered the population (thus neglecting D_{EA}) and noting $\delta_h = h_2 - h_1$, one obtains:

$$p_a^{Eq} = \frac{u}{\overline{h}\, s} + o(\xi), \qquad (4)$$

where $\overline{h} = h + 2\,\Delta h\, p\, q$ is the average dominance in the population. Note that the polymorphism at the enhancer locus increases average dominance (since $\Delta \mathrm{h} > 0$), which reduces the frequency of deleterious mutations. The direct selection term (3.a) actually stems from this effect on average dominance: fixation at the enhancer locus is similar to reducing average dominance, which is favorable. Like in models for the evolution of ploidy (38,39), direct selection favors the masking of deleterious mutations, but here this effect is frequency-dependent. We refer to this term as the 'masking' term.

The indirect selection is relatively straightforward since the expression within brackets in (3.b) can be shown to be always positive. That means that indirect selection has the same sign than D_{EA}: it is positive when the stronger enhancer (E_2) is preferentially associated with beneficial gene alleles ($D_{EA} > 0$ in this situation). The question turns now to determine the sign of D_{EA}. To do so, we use a quasi-linkage equilibrium (QLE) approximation that requires that the forces causing frequency changes are weak relative to recombination rate (75). This approximation usually breaks down at low recombination rates. However, because we are at mutation selection balance for the gene, and because the frequency change at the enhancer locus is small, an accurate approximation can be obtained by linearizing under the same assumptions as above, and keeping terms in both rD_{EA} and sD_{EA} (76). We obtain D_{EA}^{QLE}:

$$D_{EA}^{QLE} = \frac{p\, q\, p_a[2\Delta_h(1-2p) + \delta_h]s}{2r + [h_1 p^2 + 2pqh + h_2 q^2]s} + o(\xi) \qquad (5)$$

Note that this quasi-linkage equilibrium value does not diverge for small recombination rate. Noting that Δ_h is at most half as large as δ_h, it is straightforward to show that $D_{EA}^{QLE} > 0$, indicating that the indirect selection term is positive and favors stronger enhancers. This effect stems from the fact that E_2 carrying chromosomes are more exposed to selection because of their increased average expression. They are thus purged more efficiently from deleterious mutations: as a consequence E_2 alleles are most often found on, and hitchhike

with, beneficial genetic background (they are associated to *A* alleles), which is beneficial for E_2 alleles. We refer to this term as the 'purging' term.

The same model can be made with beneficial instead of deleterious mutations. In this case, we cannot assume that $p_a \ll 1$, since *a* alleles sweep to fixation. As a consequence the expression of Δp, involves more terms depending on p_a. However, Δp can still be partitioned into a 'masking' and a 'purging' term like in equations (3.a) and (3.b), respectively. Provided that beneficial mutations are dominant, and for any p_a value, the 'masking' term still favors the most frequent alleles, whereas the 'purging' term still favors stronger enhancers. Qualitatively, the model with beneficial dominant mutations and the model with deleterious recessive mutations give similar results concerning enhancer alleles' frequency dynamics. Considering partially dominant deleterious mutations would lead to a moderately different outcome (the term 3.a will switch sign, causing more enhancer polymorphism), but this scenario is biologically much less relevant.

Decrease of mean fitness

When a stronger enhancer spreads, it is favored by the purging term, but at the cost of unmasking deleterious mutations. Mean fitness equals *1 – 2u* when there is no variation at the enhancer locus (*p = 0* or *p = 1*). However, during the spread of an enhancer, mean fitness $\overline{W}$ decreases as can be readily seen from:

$$\frac{\partial \overline{W}}{\partial p} = -4\Delta h(D_{EA} + (1-2p)p_a)s\,, \qquad (6)$$

which is negative around *p < ½* and positive around *p > ½.* This is due to the fact that deleterious mutations are unmasked when average dominance increases, before they have time to reach their lower mutation – selection equilibrium frequency. A temporary genetic burden appears: deleterious mutations are too frequent given their new mean dominance.

Numerical calculation of mutant enhancer fixation probability

Because the selection coefficient on a new enhancer allele is frequency dependent (3.a), it is useful to obtain a more integrated measure of selection on enhancer alleles. One solution is to compute their probability of fixation *U* (which accounts for all frequency trajectories) and compare it to a neutral expectation. We computed it, for a mutant enhancer initially at frequency p_0, using a diffusion approximation (77):

$$U(p_0) = \int_{p=0}^{p_0} e^{-\int \frac{4N_{pop}\Delta p}{pq} dp} dp \Big/ \int_{p=0}^{1} e^{-\int \frac{4N_{pop}\Delta p}{pq} dp} dp \qquad (7)$$

As the numerical integration calculated from equation (7) relies on some assumptions, we use numerical simulations of mutant enhancer fixation or loss in a finite population to check the corresponding results. Fig 3 illustrates ratio of the probability of fixation of new enhancer alleles relative to $1/2N_{pop}$, the probability of fixation of a neutral allele. Numerical simulations were performed using a C++-program of an individual-based stochastic version of the model described above. There are N_{pop} individuals in the population. Each individual has two loci, enhancer and gene, with two alleles each. Alleles at the enhancer locus are encoded by a real value representing enhancer strength. Alleles at the gene locus carry either the wild-type allele or a deleterious allele of fitness effect *s*. The fitness of the diploid genotype is given by $w_1 - h_i (w_1 - w_2)$, were h_i is the dominance in this individual, which can vary depending on the genotype at the enhancer locus (following equation 2) and where w_1 is the fitness of the fittest of the two alleles (w_2 the fitness of the other allele). At generation 0, all individuals are homozygote for the same enhancer allele and the wild-type gene allele. Then, 2000 generations of the life cycle (diploid selection, meiosis with recombination, mutation and syngamy) are performed. During these generations, there is no mutation on the enhancer locus. At each generation, the number of mutations on the gene is sampled in a Poisson Distribution with expectation $2N_{pop}u$. A corresponding number of alleles (sampled randomly in the population) is then assigned to be deleterious. We then generate the population of N_{pop} individuals at the next generation, accounting for selection, meiosis and

random mating. For each individual in the next generation, we first determine its two parents. Two individuals of the current generation are sampled randomly. When chosen, each candidate is accepted with a probability equal to its fitness, or resampled. Once the two parents are identified, we sample one gamete in each of them (recombination occurring at a rate *r* between the two loci). After the 2000 generations, the deleterious allele frequency is close to the mutation-selection balance $u\ /\ hs$. A chromosome is then randomly chosen in the whole population and we assign it a new enhancer allele differing in strength. The new enhancer allele is then monitored until fixation or loss. Fixation probabilities were computed from 10000 runs of such simulations for each set of parameter values. The results presented on Fig 3 in the main text are obtained dividing those probabilities of fixation by that of a neutral enhancer in the same conditions (i.e. a mutant enhancer allele having the same strength than the resident allele).

Results showed on Fig 3 were obtained using the following parameters values: (1) dominance coefficient $h = 0.25$, (2) gene mutation rate $u = 10^{-3}$, (3) population size $N_{pop} = 10^3$, (4) selection coefficients *s* = 0.1 or 0.01 and (5) enhancer mutant strength three times larger or smaller than the resident allele. The *0.25* value of dominance is the most biologically plausible value for deleterious mutations (78). Results illustrated are valid for other combination of *u* and N_{pop} provided $N_{pop}u = 1$. In other situations, the results would be magnified about 1 by a factor $N_{pop}u$. For instance in a very large population $N_{pop} = 10^8$ with weaker gene mutation rate $u = 10^{-5}$, we have $N_{pop}u = 10^3$, so that tightly linked stronger enhancers would be $\sim 2.10^3$ more likely to fix than a neutral allele (instead of ~ 2 more likely as illustrated on Fig 3 for $N_{pop}u = 1$). Fig 6 illustrates this behavior, where the fixation probability of stronger enhancers, relative to the neutral expectation, scales linearly with $N_{pop}u$.

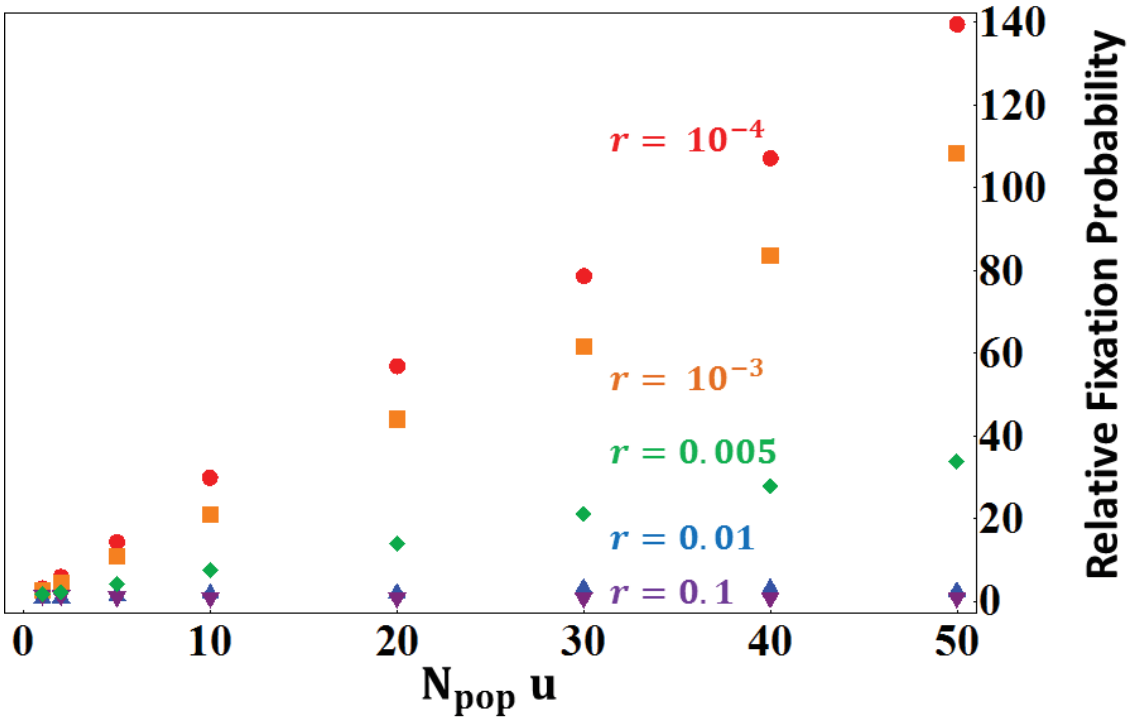


Figure 6 - Fixation probabilities ratios of stronger enhancers (3:1 strength ratio compared to the resident allele) relative to that of neutral mutation (from Eq. 7). Results were obtained for different recombination rates between the enhancer and the gene as indicated on the figure. Other parameter values are *s = 0.01*, *h*= 0.25.

Numerical simulations of enhancer strength evolution

A simplifying assumption of the model described above is to consider a population with only two enhancer alleles. We also designed an infinite-allele version of the model to check the robustness of the conclusions and to study long-term enhancer strength evolution. The goal here is to study the evolution of the mean strength of a population of enhancers undergoing recurrent unbiased mutations. In order to be consistent, we also considered the gene locus as an infinite-allele locus undergoing recurrent deleterious mutations of various effects. The simulations performed on this model are designed as the fixation probability simulations with an individual-based stochastic model, except that we do not study fixation probabilities but long-term trait evolution. At each new generation, the number of mutations on the enhancer locus is drawn from a Poisson distribution with mean $2N_{pop}u_E$, u_E being the

mutation rate on the enhancer locus per individual per generation. We denote $z_i = \mathrm{Log}(e_i)$, the logarithm of enhancer strength of allele *i*. We consider that mutations alter the trait z_i such that after mutation it becomes $z_i + \varepsilon$, where ε is a normal deviate $\varepsilon \sim N(0, \sigma_E^2)$. We considered additive mutational effect on the logarithm of enhancer strength, in order to avoid that the mutational variance vanishes on mean enhancer strength when trait value increases. Indeed, in the model, only relative enhancer strength matters in the competition for expression. As a consequence, if mean enhancer strength increases in the long run, a constant mutational variance on the trait (not its Log) would tend to produce less and less differences between enhancers such that relative enhancer strength ratios will eventually tend to 1. Mutation as described above avoids this artifact and also ensures that enhancer strength always remains positive. To model mutations on the gene, the number of mutation events is drawn for each generation from a Poisson distribution with mean $2N_{pop}u_A$, u_A being the mutation rate on the gene locus per individual per generation. For each mutation event, the fitness effect of the new mutant allele is drawn from a negative exponential distribution with mean s.

In order to measure the rate of the ER process, we follow the population mean of *z* through time. This mean increases linearly with time and we use the slope of this linear increase to compute the mean doubling time (i.e. the time needed for mean enhancer strength to double). To obtain the mean doubling time, we first store at regular time points $z_1(i,j,t)$ and $z_2(i,j,t)$, the logarithms of the strengths of both enhancers of individual *i* at generation *t* during iteration *j* (simulations are repeated 100 times). For each generation sampled, we calculate mean *z* over the whole population and over the different iterations:

$$\overline{z}(t) = \frac{1}{N_{it}} \sum_{j=1}^{N_{it}} \sum_{i=1}^{N_{pop}} \frac{z_1(i,j,t) + z_2(i,j,t)}{2N_{pop}} \qquad (8)$$

As $\overline{z}$ increase linearly with time, we estimate its rate of increase using a linear regression. Doubling times $T^{\times 2}$ is computed as:

$$T^{\times 2} = \frac{Log(2)}{a} \quad (9)$$

where a is the slope of this regression. Results of this model are illustrated on fig. 4 (red curves) and were obtained using the following parameter values: (1) dominance coefficient $h = 0.25$, (2) mutation rates $u_A = u_E = 10^{-3}$, (3) selection intensity $s = 0.1$, (4) recombination rate between the gene and its enhancer $r = 10^{-6}$, (5) initial strength of enhancers $e_0 = 1$, (6) population size $N_{pop} = 5000$, (7) number of iteration $N_{it} = 100$, (8) number of generations $N_{gen} = 100000$. Note that this model is a special case of model 2 and 3 presented below.

Models including stabilizing selection on expression levels

Model 2: Stabilizing selection on relative expression levels among genes

Here we consider a model where, for example, the proteins are enzymes implied in bio-chemical chain reactions. In this case, the dosage relationship between enzymes plays a major role in the outcome of the reactions, and so potentially on the fitness of the individuals. In model 2, the absolute amount of proteins does not have any cost on fitness, but a departure from an optimum dosage between different proteins (produced from different loci) does. In reality, dosage relationships would probably involve several proteins, but we will consider only two for simplicity and because it probably corresponds to the maximum dosage constraint. With a larger sets of, say, *n* genes in dosage relationships, the increase in enhancer strength at a focal gene will alter its dosage relationships with *n* - 1 genes. In principle, if each imbalanced gene pair contributes a cost then the overall cost should be larger for a larger set of genes. However, if, more plausibly, the set of co-regulated genes contribute to a specific biological function, then the fitness cost will be paid when this function is disrupted, but will not accumulate beyond this. Thus, it is robust to assume that the function will become more disrupted with a larger fraction of imbalanced gene pairs. Because this fraction equals 2/*n*, it is maximal for *n* = 2 (the number of imbalanced pairs is *n*-1 and the number of balanced pairs (*n*-1)(*n*-2)/2). Hence, as long as the fitness cost increases with the proportion of imbalanced genes pairs, the situation with only two genes is probably the most stringent. This model is designed similarly to model 1, but involves two major

changes. First, we need two pairs of evolving enhancer/gene chromosomes to model the expression of the two types of proteins. We denote z_{ij} the log-strength of enhancer on chromosome *i* at locus *j*. We assume a recombination rate of r between the enhancers and the corresponding genes, and a recombination rate of 0.5 between the first gene and the second enhancer (which is equivalent to saying that there are two different chromosome pairs). Second, if we note W the fitness of an individual, W_A^j the fitness of each diploid enhancer/gene set defined as previously and W_E a new fitness component taking into account the stabilizing selection on dosage relationship between the proteins, we have:

$$W = W_A^1 \times W_A^2 \times W_E \quad (10)$$

We denote Z_j the sum of the logarithmic strengths of the two alleles at the enhancer locus *j* ($Z_j = z_{1j} + z_{2j}$). We define W_E as:

$$W_E = e^{-I(Z_1 - Z_2)^2}, \quad (11)$$

where *I* stands for the intensity of stabilizing selection on the dosage relationship. We apply stabilizing selection on *Z* to avoid any bias since mutations occur on *z*. Stabilizing selection favors an optimal phenotype where the two proteins are, for simplicity, equally produced. In order to biologically scale the intensity of stabilizing selection on expression (*I*) with the impact of deleterious mutations at the protein-coding loci, we introduce a scaling parameter *γ*. Considering that fitness reduction after one round of mutations on the gene is *u h s*, and that fitness reduction after one round of mutations on regulatory sequences is $1 - e^{-I \times 4\sigma_E^2}$, we compute *I* such that:

$$1 - \gamma\, u\, h\, s = e^{-I \times 4\sigma_E^2} \quad (12)$$

With this scaling, the cost of non-optimal expression of random mutations on enhancers is approximately *γ* time the cost of recurrent deleterious mutations on the gene. We run simulations for *γ* values of *0, 1, 5* and *10*. Fig 4 illustrates these simulations with the same parameter values than for model 1.

Model 3: Stabilizing selection on absolute expression level on a focal gene

Here, we consider the second case where, contrary to model 1, enhancer polymorphism causes variation in expression levels. In this case, stabilizing selection acts directly on absolute expression levels. Any increase or decrease from the optimal amount of proteins is deleterious. In this situation, we also allow for trans-acting regulatory factors (referred to as transcription factors below, TFs) to evolve as well, as they may compensate for the evolution of stronger enhancers. In model 3, the absolute level of protein expression depends on the strengths of enhancers and TFs. The goal here is to investigate if enhancer strength evolution is limited by the constraint on absolute expression levels or if an open ended coevolution can take place between enhancers, whose strength tends to increase due to the ER process, and the TFs, for which we predict that they should coevolve such as to maintain optimal protein expression. For this model we designed simulations with three loci: the TF locus, the enhancer locus and the gene locus. TF locus is not localized on the same chromosome, i.e. it recombines freely with the two other loci, $r_{TE} = 0.5$. It influences expression on the gene in *trans*, i.e. each allele at this locus interacts with the two homologous enhancer alleles. Following the same notation as previously, we note Z_1 and Z_2 the logarithmic strengths of enhancer and TF loci, respectively. Without loss of generality, we assume that optimal amount of proteins is produced when $Z_1 + Z_2 = 0$ (meaning that the overall strength of TFs exactly opposes the overall strength of enhancers):

$$\begin{cases} W = W_A \times W_E \\ W_E = e^{-I(Z_1+Z_2)^2}, \end{cases} \quad (13)$$

The intensity of the stabilizing selection *I* is scaled as in the previous model. Like for the enhancer locus, the number of mutations on the TF locus is drawn from a Poisson distribution with mean $2N_{pop}u_T$; and these mutations additively change the Log of the trait.

Results on Fig 4 are presented in terms of doubling time as previously and were obtained with the same parameter values than before.

Effect of inbreeding on the ER process (model 4)

In model 4, we want to study modifications in escalation rates resulting from variation in the mating system. Here, we use the same assumptions than in model 1 except that each individual gets a probability p_s to self-fertilize. Self-fertilization is modelled by sampling the second gamete from the same diploid parent than the first. Doubling times are calculated as previously and are reported on Fig 5.

Acknowledgments

We thank Marie-Claude Quidoz, and the CEFE computing platform, as well as the MBB computing platform for helping with the cluster resources. We thank Giacomo Cavalli and Bernard de Massy for useful discussions and Christoph Haag and Luis-Miguel Chevin for useful comments on the manuscript.

References

1. Hoekstra HE, Coyne JA. The locus of evolution: evo devo and the genetics of adaptation. Evolution (N Y). 2007;61(5):995–1016.

2. Carroll SB. Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell. 2008;134(1):25–36.

3. King M, Wilson A. Evolution at two levels in humans and chimpanzees. Science (80-). 1975;188(4184):107–16.

4. Abzhanov A, Protas M, Grant BR, Grant PR, Tabin CJ. Bmp4 and morphological variation of beaks in Darwin's finches. Science. 2004;305(5689):1462–5.

5. Wilson AC, Maxson LR, Sarich VM. Two types of molecular evolution. Evidence from studies of interspecific hybridization. Proc Natl Acad Sci U S A. 1974;71(7):2843–7.

6. Cooper T. Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. Proc Natl Acad Sci U S A. 2003;100(3):1072–7.

7. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman M V, et al. The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol. 2003;20(9):1377–419.

8. Fay JC, Wittkopp PJ. Evaluating the role of natural selection in the evolution of gene regulation. Heredity (Edinb). 2008;100(2):191–9.

9. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat Rev Genet. 2012;13(1):59–69.

10. Pastinen T, Hudson TJ. Cis-acting regulatory variation in the human genome. Science. 2004;306(5696):647–50.

11. Cheung VG, Conlin LK, Weber TM, Arcaro M, Jen K-Y, Morley M, et al. Natural variation in human gene expression assessed in lymphoblastoid cells. Nat Genet. 2003;33(3):422–5.

12. Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010;464(7289):768–72.

13. Ferea TL, Botstein D, Brown PO, Rosenzweig RF. Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc Natl Acad Sci U S A. 1999;96:9721–6.

14. Gompel N, Prud'homme B, Wittkopp PJ, Kassner VA, Carroll SB. Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. Nature. 2005;433(7025):481–7.

15. Raymond M, Chevillon C, Guillemaud T, Lenormand T, Pasteur N. An overview of the evolution of overproduced esterases in the mosquito Culex pipiens. Philos Trans R Soc B Biol Sci. 1998;353(1376):1707–11.

16. Shapiro MD, Marks ME, Peichel CL, Blackman BK, Nereng KS, Jónsson B, et al. Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature. 2004;428(6984):717–23.

17. Wagner GP, Lynch VJ. The gene regulatory logic of transcription factor evolution. Trends Ecol Evol. 2008;23(7):377–85.

18. Jenkins DJ, Stekel DJ. De Novo Evolution of Complex, Global and Hierarchical Gene Regulatory Mechanisms. J Mol Evol. 2010;71(2):128–40.

19. Jenkins DJ, Stekel DJ. A New Model for Investigating the Evolution of Transcription Control Networks. Artif Life. 2009;15(3):259–91.

20. Crombach A, Hogeweg P. Evolution of Evolvability in Gene Regulatory Networks. PLoS Comput Biol. 2008;4(7):e1000112.

21. Aldana M, Balleza E, Kauffman S, Resendiz O. Robustness and evolvability in genetic regulatory networks. J Theor Biol. 2007;245(3):433–48.

22. Quayle AP, Bullock S. Modelling the evolution of genetic regulatory networks. J Theor Biol. 2006;238(4):737–53.

23. Ptashne M, Gann A. Transcriptional activation by recruitment. Nature. 1997;386(6625):569–77.

24. Spitz F, Furlong EEM. Transcription factors: from enhancer binding to developmental control. Nat Rev Genet. 2012;13(9):613–26.

25. Villar D, Flicek P, Odom DT. Evolution of transcription factor binding in metazoans [mdash] mechanisms and functional implications. Nat Rev Genet. 2014;15(4):221–33.

26. Rockman M V, Wray GA. Abundant raw material for cis-regulatory evolution in humans. Mol Biol Evol. 2002;19(11):1991–2004.

27. Knight JC. Allele-specific gene expression uncovered. Trends Genet. 2004;20(3):113–6.

28. Lo HS, Wang Z, Hu Y, Yang HH, Gere S, Buetow KH, et al. Allelic variation in gene expression is common in the human genome. Genome Res. 2003;13(8):1855–62.

29. Whitehead A, Crawford DL. Neutral and adaptive variation in gene expression. Proc Natl Acad Sci. 2006;103(14):5425–30.

30. Khaitovich P, Weiss G, Lachmann M, Hellmann I, Enard W, Muetzel B, et al. A Neutral Model of Transcriptome Evolution. PLoS Biol. 2004;2(5):e132.

31. Denver DR, Morris K, Streelman JT, Kim SK, Lynch M, Thomas WK. The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat Genet. 2005;37(5):544–8.

32. Gilad Y, Oshlack A, Smyth GK, Speed TP, White KP. Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature. 2006;440(7081):242–5.

33. Ludwig MZ, Bergman C, Patel NH, Kreitman M. Evidence for stabilizing selection in a eukaryotic enhancer element. Nature. 2000;403(6769):564–7.

34. Lemos B, Meiklejohn CD, Cáceres M, Hartl DL. Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability among functional categories. Evolution. 2005;59(1):126–37.

35. Rifkin SA, Kim J, White KP. Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet. 2003;33(2):138–44.

36. Whitehead A, Crawford DL. Variation within and among species in gene expression: raw material for evolution. Mol Ecol. 2006;15(5):1197–211.

37. Manna F, Martin G, Lenormand T. Fitness Landscapes: An Alternative Theory for the Dominance of Mutation. Genetics. 2011;189(3):923–37.

38. Otto SP, Goldstein DB. Recombination and the evolution of diploidy. Genetics. 1992;131:745–51.

39. Cailleau A, Cheptou P-O, Lenormand T. Ploidy and the evolution of endosperm of flowering plants. Genetics. 2010;184(2):439–53.

40. Fisher RA. The Evolution of Dominance. Biol Rev. 1931;6(4):345–68.

41. Austin B, Trivers R, Burt A. Genes in conflict: the biology of selfish genetic elements. Harvard University Press; 2009.

42. Kirkpatrick M. Sexual selection and the evolution of female choice. Evolution (N Y). JSTOR; 1982;1–12.

43. Cheung VG, Nayak RR, Wang IX, Elwyn S, Cousins SM, Morley M, et al. Polymorphic Cis- and Trans-Regulation of Human Gene Expression. PLoS Biol. 2010;8(9):e1000480.

44. Kuo D, Licon K, Bandyopadhyay S, Chuang R, Luo C, Catalana J, et al. Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Res. 2010;1–7.

45. Proulx SR, Phillips PC. The Opportunity for Canalization and the Evolution of Genetic Networks. Am Nat. 2005;165(2):147–62.

46. Wagner GP, Bürger R. On the evolution of dominance modifiers II: a non-equilibrium approach to the evolution of genetic systems. J Theor Biol. 1985;113(3):475–500.

47. Otto SP, Yong P. The evolution of gene duplicates. Homol Eff. 2002;46:451–83.

48. Otto SP, Bourguet D. Balanced Polymorphisms and the Evolution of Dominance. Am Nat. 1999;153(6):561–74.

49. Bourguet D. The evolution of dominance. Heredity (Edinb). 1999;83(1):1–4.

50. Wright S. Fisher's Theory of Dominance. Am Nat. 1929;63(686):274–9.

51. Masel J, Siegal ML. Robustness: mechanisms and consequences. Trends Genet. 2009;25(9):395–403.

52. Denby CM, Im JH, Yu RC, Pesce CG, Brem RB. Negative feedback confers mutational robustness in yeast transcription factor regulation. Proc Natl Acad Sci. 2012;109(10):3874–8.

53. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. BioEssays. 1998;20(5):433–40.

54. Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, et al. A genome-wide association study of global gene expression. Nat Genet. 2007;39(10):1202–7.

55. Lenski RE, Ofria C, Pennock RT, Adami C. The evolutionary origin of complex features. Nature. 2003;423(6936):139–44.

56. Lynch M. The frailty of adaptive hypotheses for the origins of organismal complexity. Proc Natl Acad Sci. 2007;104:8597–604.

57. Soyer OS, Bonhoeffer S. Evolution of complexity in signaling pathways. Proc Natl Acad Sci. 2006;103(44):16337–42.

58. Biggin MD. Animal Transcription Networks as Highly Connected, Quantitative Continua. Dev Cell. 2014;21(4):611–26.

59. Ochoa-Espinosa A, Yucel G, Kaplan L, Pare A, Pura N, Oberstein A, et al. The role of binding site cluster strength in Bicoid-dependent patterning in Drosophila. Proc Natl Acad Sci U S A. 2005;102(14):4960–5.

60. Lenormand T, Roze D, Rousset F. Stochasticity in evolution. Trends Ecol Evol. 2015;24(3):157–65.

61. Weirauch MT, Hughes TR. Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. 2010;26(2):66–74.

62. Yvert G, Brem RB, Whittle J, Akey JM, Foss E, Smith EN, et al. Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet. 2003;35(1):57–64.

63. Tirosh I, Reikhav S, Levy AA, Barkai N. A yeast hybrid provides insight into the evolution of gene expression regulation. Science. 2009;324(5927):659–62.

64. Wittkopp PJ, Haerum BK, Clark AG. Regulatory changes underlying expression differences within and between Drosophila species. Nat Genet. 2008;40(3):346–50.

65. Cowles CR, Hirschhorn JN, Altshuler D, Lander ES. Detection of regulatory variation in mouse genes. Nat Genet. 2002;32(3):432–7.

66. Wittkopp PJ, Haerum BK, Clark AG. Evolutionary changes in cis and trans gene regulation. Nature. 2004;430(6995):85–8.

67. He F, Zhang X, Hu J, Turck F, Dong X, Goebel U, et al. Genome-wide Analysis of Cis-regulatory Divergence between Species in the Arabidopsis Genus. Mol Biol Evol. 2012;29(11):3385–95.

68. Steige KA, Reimegård J, Koenig D, Scofield DG, Slotte T. Cis-Regulatory Changes Associated with a Recent Mating System Shift and Floral Adaptation in Capsella. Mol Biol Evol. 2015;

69. Mellert DJ, Truman JW. Transvection Is Common Throughout the Drosophila Genome. Genetics. 2012;191(4):1129–41.

70. Heride C, Ricoul M, Kiêu K, von Hase J, Guillemot V, Cremer C, et al. Distance between homologous chromosomes results from chromosome positioning constraints. J Cell Sci. 2010;123(23):4063–75.

71. Rieder D, Trajanoski Z, McNally JG. Transcription factories. Front Genet. 2012;3:221.

72. Gerton JL, DeRisi J, Shroff R, Lichten M, Brown PO, Petes TD. Global mapping of meiotic recombination hotspots and coldspots in the yeast Saccharomyces cerevisiae. Proc Natl Acad Sci. 2000;97(21):11383–90.

73. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova G V. Genetic recombination is directed away from functional genomic elements in mice. Nature. 2012;485(7400):642–5.

74. Sen R, Grosschedl R. Memories of lost enhancers. Genes Dev. 2010;24(10):973–9.

75. Nagylaki T. The evolution of multilocus systems under weak selection. Genetics. 1993;134(2):627–47.

76. Roze D. Selection for sex in finite populations. J Evol Biol. 2014;27(7):1304–22.

77. Kimura M. Diffusion Models in Population Genetics. J Appl Stat. 1964;1(2):177–232.

78. Manna F, Gallet R, Martin G, Lenormand T. The high-throughput yeast deletion fitness data and the theories of dominance. J Evol Biol. 2012;25(5):892–903.

La modélisation formelle d'un locus gène sous sélection, associé à un locus *cis*-régulateur polymorphe, nous a permis de nous rendre compte qu'il pouvait exister une pression de sélection favorisant les promoteurs plus forts. La raison est que, si la recombinaison entre les deux locus est suffisamment faible, des associations génétiques se créent entre promoteurs forts et allèles viables, ce qui permet aux promoteurs forts d'envahir la population par auto-stop.

On a vu que ce processus dépend de quatre paramètres majeurs : l'intensité de la sélection sur le gène, le taux de mutation du gène, le taux de recombinaison entre le gène et le promoteur et l'intensité de la sélection stabilisante sur les niveaux d'expression. La sélection stabilisante a tendance à ralentir l'augmentation de la force des promoteurs. La force des promoteurs augmente tout de même, car des mutations compensatoires ont lieu dans d'autres régulateurs pour assurer des niveaux d'expression optimaux.

Afin de mieux comprendre cette coévolution entre régulateurs, nous avons étendu le modèle. Nous étudions, dans la suite, les possibilités de coévolution entre un *cis*- et un *trans*-régulateur, entre deux *cis*-régulateurs du même gène, et entre un *cis*-régulateur et une boucle de rétroaction négative. Ces modèles permettent de montrer que les coévolutions sont possibles avec tout type de régulateurs, à condition que la sélection stabilisante sur les niveaux d'expression ne soit pas trop forte, et tolère temporairement des niveaux d'expression sous-optimaux. En utilisant un autre type de modèle (voir Annexe), nous mettons en évidence que l'*ER process* peut mener à des réseaux de régulation inutilement complexes, faisant intervenir de nombreux acteurs et de nombreuses interactions quand une architecture plus simple pourrait assurer des niveaux d'expression optimaux. On montre également que l'*ER process* peut entraîner un gradient dans la force des promoteurs à mesure que ceux-ci s'éloignent de leur gène.

Nous avons également étudié avec un peu plus de précision les conséquences de l'*ER process* sur les coefficients de dominance des allèles délétères du gène. Nous nous sommes rendus compte qu'il existe une petite fenêtre de recombinaison au sein de laquelle les promoteurs forts envahissent bien qu'ils augmentent la dominance moyenne des allèles délétères, ce qui est défavorable. Quand la recombinaison est plus faible, les promoteurs forts envahissent tout en diminuant la dominance moyenne ; quand la recombinaison est

plus forte, les promoteurs forts n'envahissent plus. Dans cette fenêtre, il y a la possibilité qu'un suppresseur d'expression allele-specifique (*ASE*) diminuant les écarts d'expression entre allèle homologues envahisse. Un tel suppresseur diminuera la dominance moyenne et rendra plus difficile l'invasion de promoteurs mutants, puisque cette invasion repose justement sur des écarts d'expression entre allèles homologues. Dans cette petite fenêtre, les promoteurs forts se comportent comme des éléments égoïstes, et un suppresseur évolue : un conflit génétique a lieu.

Paper in preparation

Chapitre II

ER process-driven Coevolution in Regulatory Networks of Diploids

Authors : Frédéric Fyon, Thomas Lenormand

UMR 5175 CEFE, CNRS - Université Montpellier - Université P. Valéry - EPHE, 1919 route de Mende 34293 Montpellier Cedex 5, France

Introduction

The evolution of regulatory networks present distinctive features compared to evolution on protein-coding genes: (1) gene regulation plays a prominent role in organismal development, which means that disruption of expression profiles are mostly selected against (stabilizing selection seems to apply throughout regulatory networks) (LUDWIG *et al.* 2000; FAY and WITTKOPP 2008); (2) different regulatory networks may be functionally equivalent and lead to similar expression profiles and phenotypes (WEIRAUCH and HUGHES 2010). This combination of widespread stabilizing selection and neutral equivalence among different networks can result in distinctive evolutionary patterns: regulatory sequences may often change, while maintaining constant expression profiles. With such a mode of evolution, phenotypic distance can be poorly correlated to genetic distance (see WRAY *et al.* (2003) and references therein). Almost similar networks may result in very different expression profiles, whereas different networks may result in similar expression profiles (TAUTZ 2000). This allows for mutations slightly disrupting expression profiles to be maintained and compensated by other mutations restoring expression profiles (KUO *et al.* 2010). This feature of regulatory networks results in evolution to run idle: evolutionary changes occur through genetic drift, but expression profiles are maintained (WEIRAUCH and HUGHES 2010). In this paper, we argue that this phenomenon may be accelerated by a positive indirect selective process on regulatory networks.

In a previous paper (FYON *et al.* 2015), we used population genetics equation and individual-based multi-locus models to identify a selective process favoring the open-ended evolution of stronger enhancers, when they are sufficiently close to the gene they regulate. This escalation is only expected to be stopped by physical limitation. We referred to it as the '*Enhancer Runaway*' (ER) process. ER process is based on allele-specific expression in cases of *cis*-regulatory polymorphism. *Cis*-acting regulatory sequences, which include enhancers and promoters, control expression of genes specifically located on the same chromosome (WRAY *et al.* 2003). In *cis*-regulatory heterozygotes, a gene allele associated with a stronger enhancer (an enhancer allele that activates more transcription initiation), will be more expressed than the homologous allele, associated with a weaker enhancer. As a result, it will be more exposed to selection. This means that a stronger enhancer will better purge its

associated gene copy from deleterious mutations. Hence, stronger enhancers become associated to better genetic background, and increase in frequency by indirect selection.

In this chapter, we investigate the multiple possibilities for coevolution within regulatory networks triggered by this process. Indeed, this process may lead to two types of costs. First, when a stronger enhancer is invading a population, mean dominance of deleterious alleles can be temporarily increased, which is detrimental at the individual level. This means that a global suppressor of ER process could potentially be selected for, to prevent this increase in mean dominance. Second, the ER process can lead to protein over-expression. If expression levels are under stabilizing selection, this over-expression is costly: it may disrupt metabolic processes, alter cytoplasmic viscosity and consume unnecessarily amounts of energy and resources. As a response to this over-expression, various types of other regulatory elements may evolve to compensate and restore optimal expression levels. Here, we investigate possible invasion of three types of regulators: *trans*-acting regulators, *cis*-acting regulators and *trans*-acting feedback loop.

Overall, we argue here that the ER process does not actually concern only *cis*-regulatory sequences in a region close from the gene. Due to its impact on expression levels and allelic dominance, it can have a profound and widespread impact on the evolution of eukaryotic regulation networks.

Methods

Allele specific expression (ASE) suppressor

To understand the effect of enhancer polymorphism on the dominance coefficient of gene deleterious alleles, we first consider a standard deterministic population genetics model. We assume an infinite population. Each individual has three loci: an expression modifier locus **M**, an enhancer locus **E** and a gene locus **A**. Each locus has two alleles. The gene locus has a viable allele *A* with fitness *1*, and a deleterious allele *a* with fitness *1 – s*. Heterozygote individuals at the gene locus have a fitness *1 – h s*, with *h* the default dominance coefficient of gene deleterious allele. The enhancer locus has a stronger allele *E* with strength $\mathbf{e_1}$ and a weaker allele *e* with strength $\mathbf{e_2 < e_1}$. As explained below, the strength of the enhancer alleles determines the share of proteins expressed from the two homologous gene copies. The modifier locus has a wild type allele *m* that does not change expression and dominance, and a mutant allele *M* with trait **m**. We assume for simplicity that enhancer strength does not change the total amount of proteins produced. Enhancer strength only modifies the relative expression of gene alleles in gene heterozygotes. For example, in a *EA/ea* individual, *A* allele is expressed in proportion $\mathbf{e_1/(e_1+e_2) > 0.5}$, while *a* allele is expressed in proportion $\mathbf{e_2/(e_1+e_2) < 0.5}$. This is actually similar to changing the dominance coefficient of the gene deleterious allele. Indeed, in the example above, *a* allele is less expressed than in a standard heterozygote: it is thus necessarily more recessive. We define h_1 and h_2 the dominance of the deleterious allele in genotypes *EA/ea* and *Ea/eA* respectively. h_1 and h_2 have to follow a particular relationship between the fitness of an individual *W* and the proportion of gene deleterious allele expression *%a*. This relationship must be decreasing, most obviously, and concave, as deleterious alleles are on average recessive. The fitness effect of a deleterious mutation must increase more than linearly with its proportion of expression, and must correspond to *hs* when it is expressed at 50%. We use the following relationship:

$$W = 1 - (\%a)^{-\frac{Log(h)}{Log(2)}s} \quad (1)$$

As a consequence, we have:

$$\begin{cases} h_1 = \left(\frac{e_2}{e_1 + e_2}\right)^{-\frac{Log(h)}{Log(2)}} < h & \text{(2.a)} \\ h_2 = \left(\frac{e_1}{e_1 + e_2}\right)^{-\frac{Log(h)}{Log(2)}} > h & \text{(2.b)} \end{cases}$$

We consider that these h_1 and h_2 dominance coefficients may be changed by the modifier locus. Its effect is to buffer allele-specific expression. Biologically, such a modifier could correspond to a mechanism limiting competition for transcription factors between *cis*-regulatory elements. For instance, in mammals, the territories of homologous chromosomes within the nucleus tend to be more distant compared to those of a random pair of autosomes (HERIDE *et al.* 2010). This spatial partitioning can strongly limit competition for transcription factors, and thus ASE patterns, as only factors available locally will interact with each *cis*-regulatory element. Inversely, chromosome pairing and transvection in *Diptera* tend to produce balanced expression between homologous chromosomes (MELLERT and TRUMAN 2012). We model this kind of mechanism by considering that the expression modifier tends to alter dominance coefficients h_1 and h_2 so that they become closer to h: it is an ASE suppressor. Effects at this locus are considered additive. If we note m the total modifier trait, then $m = 0$ for *mm* individuals, $m = \Delta m$ for *Mm* individuals and $m = 2\Delta m$ for *MM* individuals. Noting h_i and h_i' the dominance without and with expression modifier effects taken into account, we have

$$h_i' = h + \frac{h_i - h}{1 + m} \quad \text{(3)}$$

Recombination occurs between the modifier and enhancer loci at rate R_{ME} and between enhancer and gene loci at rate R_{EA}. Finally, we consider that recurrent deleterious mutations occur on the gene, at a rate u. The following equations are obtained at the leading order

using a set of approximations. Considering a small parameter ξ, we assume small selection intensity (s of the order of ξ), very small mutation rates (u of the order of ξ^2). As a consequence, deleterious mutations reach a small frequency p_a at equilibrium (p_a of the order of ξ). There is no reason for the expression modifier to be close to the enhancer and gene loci. We thus suppose that the recombination rate between modifier and enhancer loci R_{ME} is large enough for the linkage disequilibrium involving the modifier locus to be very small (D_{MEA} and D_{ME} of the order of ξ^2). As we explained in a previous paper (FYON *et al.* 2015), the ER process occurs only for tightly linked enhancers, which exhibit important linkage disequilibrium with the gene. We thus suppose that the linkage disequilibrium between enhancers and gene alleles is only small (D_{EA} of the order of ξ).

Stochastic simulations with ASE suppressors

In order to obtain results on the invasion of ASE suppressors in a more dynamic and in a stochastic version of the model, we run individual-based simulations of the same model. Here, population is finite. Genomes contain one gene locus, one enhancer locus and one expression modifier locus, all of which having an infinite number of possible alleles. Individuals go through the same life cycle than before: selection, meiosis with recombination, mutation, and syngamy. Noting $w_{i,1}$ and $w_{i,2}$ the fitness of both gene copies of an individual such that $w_{i,2} > w_{i,1}$, we define the fitness of individual i as:

$$W_i = w_{i,2} - h_i\left(w_{i,2} - w_{i,1}\right) \tag{4}$$

h_i is the dominance of gene allele of fitness $w_{i,2}$ in individual i. It is calculated as:

$$h_i = h - \frac{\left(\left(\frac{e_{i,2}}{e_{i,1}+e_{i,2}}\right)^{-\frac{Log(h)}{Log(2)}} - h\right)}{1+\frac{m_{i,1}+m_{i,2}}{2}}, \tag{5}$$

with h being the intrinsic dominance coefficient of deleterious alleles, set at $h = 0.25$ through all simulations, which is the consensus average value of the dominance of mildly deleterious mutations (MANNA *et al.* 2011). $e_{i,1}$ and $e_{i,2}$ are the strengths of enhancer alleles associated with gene alleles of fitness $w_{i,1}$ and $w_{i,2}$ respectively. $m_{i,1}$ and $m_{i,2}$ are the trait values of expression modifier alleles in individual i. Selection is simulated by sampling individuals as parents of a descendent individual with probability equal to their fitness. A chromosome of the parent individual is then chosen randomly, after recombination and mutation. Mutations occur following Poisson laws with mean $N_{pop}u_A$, $N_{pop}u_E$, $N_{pop}u_M$, for gene, enhancer and expression modifier loci respectively. u_A, u_E, u_M are the mutation rates per individual per generation for the same three loci respectively. When a mutation occurs on the gene locus, the mutant fitness effect is drawn from a negative exponential distribution with mean $1/s$. When a mutation occurs on the enhancer locus, enhancer log-strength increases (or decreases) additively by a factor drawn from a normal distribution with mean 0 and variance σ_E^2. We consider here additive mutations on the logarithm of the strength to insure that mutational variance does not vanish over enhancer mean and to insure strictly positive enhancer strength (see FYON *et al.* 2015). Similarly, when a mutation occurs on the expression modifier, the modifier trait log-value increases or decreases addititvely by a factor drawn from a normal distribution with mean 0 and variance σ_M^2.

During the first 1000 generations, only mutations on the gene occur ($u_E = u_M = 0$). The gene locus reaches the mutation-selection equilibrium. Then, for 50000 generations, only mutations on the gene and enhancer loci occur ($u_M = 0$). The ER process occurs then as described in FYON *et al.* (2015). Finally, for all subsequent generations, all loci undergo mutations. We record mean dominance and trait values for the three loci at regular time steps.

Stochastic simulations with stabilizing selection and regulatory coevolution

In a second set of models, we want to investigate the possibility of invasion of different regulators in cases of stabilizing selection on the total amount of protein produced. The model is basically the same as above, except that the modifier locus corresponds to three types of regulators: another enhancer locus (model A), a transcription factor (TF) locus (model B) or a *trans*-acting regulator involved in a negative feedback regulatory loop (model C). The life cycle is identical. Fitness of individuals is computed as the product of two effects. The first effect corresponds to the impact of deleterious mutations on the gene W_A and the second to the impact of a departure from optimal level of expression (caused by stabilizing selection on expression levels) W_s. W_A is defined as W_i in Eq. 4, except that h_i is now defined as $\left(\frac{e_{i,2}}{e_{i,1}+e_{i,2}}\right)^{-\frac{Log(h)}{Log(2)}}$. W_s is defined as:

$$W_s = e^{-I(Q-Q^*)^2}, \tag{6}$$

with Q and Q^* respectively the actual and optimal amounts of proteins produced and I the stabilizing selection intensity. In models A, B and C, the regulator has a specific way to influence the amount of proteins produced Q. In model A, Q is determined by adding log-forces of both enhancer loci. If $m > 1$, the modifier increases expression levels, if $m < 1$ it decreases expression levels. In model B, enhancer log-forces are multiplied by the trait value of the modifier. Expression levels increase if $m > 1$ and decrease if $m < 1$. In model C, the modifier brings expression levels closer or away from the optimum. If $m > 1$, expression levels depart from the optimum. If $m < 1$, expression levels get closer to the optimum. We calculate expression levels as:

$$\begin{cases} Q = Log(m_1) + Log(m_2) + Log(e_1) + Log(e_2) & \text{(7.A)} \\ Q = m_1 m_2 \big(Log(e_1) + Log(e_2)\big) & \text{(7.B)} \\ Q = Q^* + m_1 m_2 (Log(e_1) + Log(e_2) - Q^*) & \text{(7.C)} \end{cases}$$

Throughout the simulations, we will assume $Q^* = 2$, $e_1(t=0) = e_2(t=0) = 10$ and $m_1(t=0) = m_2(t=0) = 1$. Mutations are simulated as above. As previously, populations undergo no mutation at the enhancer and modifier loci during the first 1 000 generations. For the next 200 000 generations, there is still no mutation on the modifier locus, but mutations on the enhancer locus occur. Finally, mutations occur on all three loci for the remaining part of the simulations (last *200 000* generations). At regular time steps, we record mean trait values for the three loci.

Results

Analysis without any modifier

Using the assumptions presented in the methods section and assuming there is no expression modifier whatsoever, the mean fitness of the individuals in the population $\overline{W}$ is, at leading order:

$$\overline{W} = 1 - 2u\left(\frac{2\bar{h}}{h} - 1\right) \quad (8)$$

In Eq. 8, $\bar{h}$ stands for the mean dominance of the deleterious allele in the population. When the population is monomorphic for the enhancer locus, $\bar{h} = h$, which corresponds to the usual result that $\overline{W} = 1 - 2u$. Eq. 8 also indicates that when polymorphism at the enhancer locus leads to $\bar{h} > h$ (increased mean dominance), the mean fitness of the population is reduced below *1 – 2u*. Conversely, if $\bar{h} < h$ (decreased mean dominance), the mean fitness of the population increases above *1 – 2u*.

In FYON *et al.* (2015), we showed that, at leading order, $\bar{h} = h + 2\Delta h\, p\,(1-p)$, with *p* the frequency of the stronger enhancer allele and $\Delta h = \frac{(h_1+h_2)}{2} - h > 0$. We thus argued that enhancer polymorphism would always lead to increased mean dominance and decreased mean fitness. This is actually not always the case when the effect of linkage disequilibrium cannot be neglected. A more accurate result is:

$$\bar{h} = h + 2\Delta h\, p\,(1-p) - \frac{D_{EA}}{p_a}(h_2 - h - 2\Delta h\, p) \quad (9)$$

As $2\Delta h = h_1 + h_2 - 2h$, one can see that $h_2 - h = 2\Delta h + h - h_1$. Since $h > h_1$, it is clear that $h_2 - h > 2\Delta h$. Thus, we see that a positive linkage disequilibrium D_{EA} (which occurs when a stronger enhancer allele is sweeping, see FYON *et al.* 2015) tends to decrease the mean dominance in the population. This is due to the fact that a positive linkage disequilibrium corresponds to a situation with an excess of *EA/ea* genotypes among the

double heterozygotes, which has a dominance coefficient of $h_1 < h$. The effect of enhancer polymorphism on mean dominance depends thus on the magnitude of D_{EA}: it tends to increase it if D_{EA} is low enough (effect of Δh in the second term of Eq. 9) and to decrease it if D_{EA} is high enough (third term of Eq. 9).

The invasion of a stronger enhancer also depends on the linkage disequilibrium D_{EA}. From FYON *et al.* (2015), the frequency change after one generation of the stronger enhancer allele Δp is:

$$\Delta p = -2\Delta h\, p_a\, p\,(1-p)(1-2p)s + D_{EA}s\big(h_2 - (h_2 - h_1)p - 4\Delta h\, p\,(1-p)\big) \qquad (10)$$

It is possible to show that $h_2 - (h_2 - h_1)p - 4\Delta h\, p\,(1-p) > h_1 - 4\Delta h > 0$. As explained in FYON *et al.* (2015), positive linkage disequilibrium between stronger enhancer allele and viable gene allele selectively favors for the spread of the stronger enhancer allele. Interestingly, Eq. 10 can be written as:

$$\Delta p = -p_a s(1-2p)\big(\bar{h} - h\big) - D_{EA}s(1-4p) \qquad (11)$$

Eq. 11 shows that the spread of stronger enhancer alleles does not only depend on its effect on mean dominance, but also on the magnitude of the linkage disequilibrium. Linkage magnitude is notably determined by the relative strength of enhancer alleles, the intensity of the purifying selection on the gene and the recombination rate between the gene and the enhancer. With strong linkage disequilibrium, stronger enhancers spread and decrease mean dominance ($\bar{h} < h$), which increases population mean fitness. At low linkage disequilibrium, stronger enhancers cannot invade, and mean dominance is increased by regulatory mutations failing to spread ($\bar{h} > h$). At intermediate linkage, there is a window where stronger enhancers spread despite increasing mean dominance ($\bar{h} > h$), thus reducing population mean fitness. This is because the effect of the linkage disequilibrium (second term in Eq. 11) overcome the effect of increased mean dominance (first term in Eq. 11). In this window, a stronger enhancer can be viewed as being “selfish”: stronger enhancer alleles

get preferentially associated with beneficial genetic background despite causing on average a dominance cost at the individual level.

The evolution of ASE suppressor

We consider a modifier locus suppressing the ER process by buffering differences between h_1 and h_2. The frequency change over one generation of a suppressor allele with effect Δm is:

$$\Delta p_M = \frac{2\Delta m\left(1+\Delta m(1-p_M)\right)}{(1+\Delta m)(1+2\Delta m)} p_M(1-p_M)[2\Delta hp(1-p)p_a + D_{EA}(h-h_2+2\Delta hp)]s + o(\xi^2), \quad (12)$$

If we assume that expression modifiers do not alter much deleterious allele frequency, that is $p_a \approx p_a(m=0)$, Eq. 12 simplifies to:

$$\Delta p_M = \frac{2\Delta m\left(1+\Delta m(1-p_M)\right)}{(1+\Delta m)(1+2\Delta m)} p_M(1-p_M)p_a s\left(\bar{h}-h\right) \quad (13)$$

From Eq. 13, it is straightforward to see that the modifier will spread only if $\bar{h} > h$. In such cases, the modifier will cause a decrease of mean dominance of deleterious alleles ($\bar{h}$ is becoming closer to *h*), which increases population mean fitness.

Stochastic simulations with ASE suppressor

To confirm our analytical results, we ran simulations of an individual-based version of our three-locus model. We tracked mean enhancer strength, mean expression modifier trait value and mean dominance of deleterious alleles over generations. They are illustrated on Fig. 1.

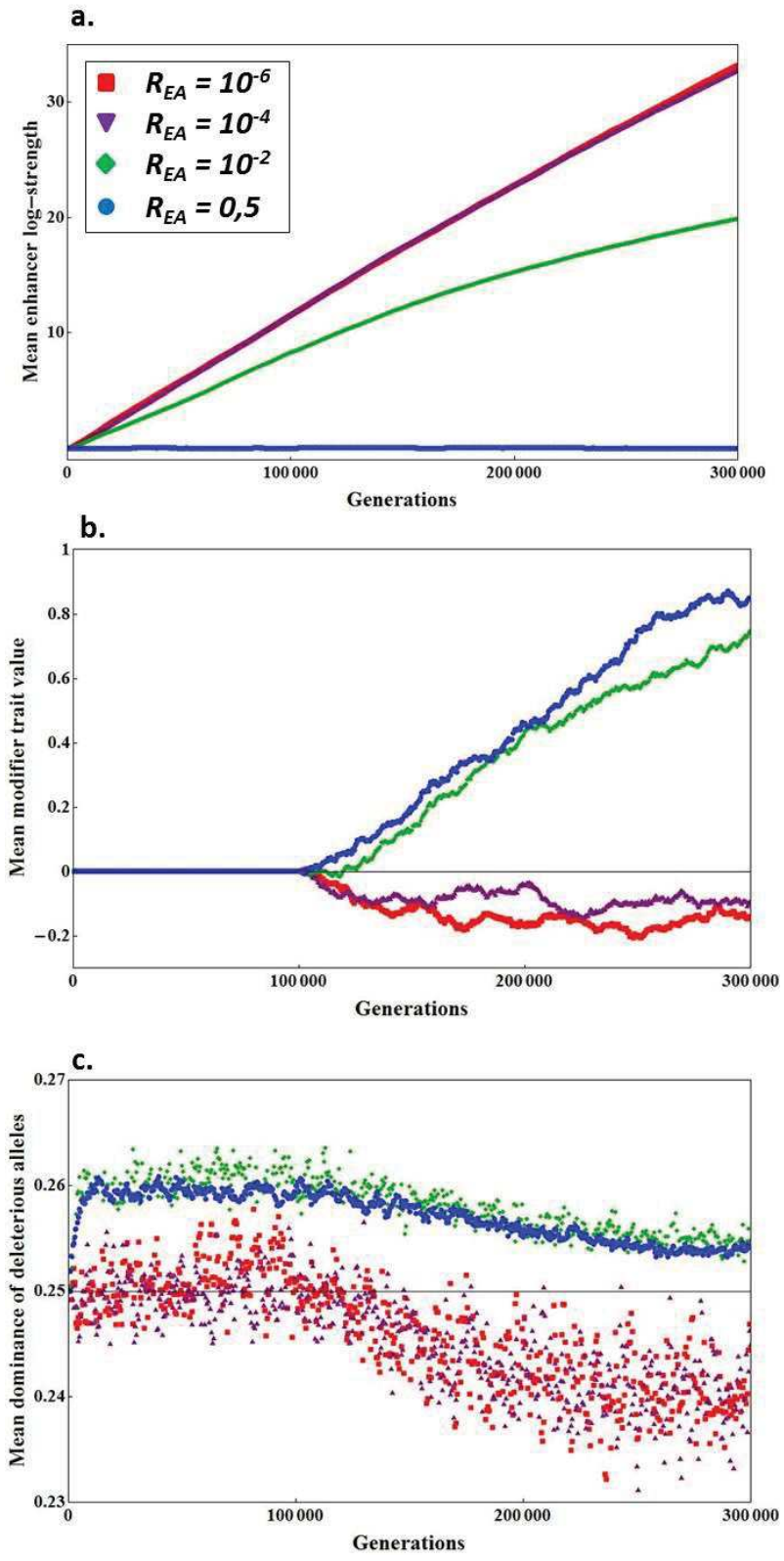


Figure 1. Simulated evolution of one gene, one enhancer and one ASE suppressor over 300 000 generations. We follow mean enhancer log-strength (**a**), mean modifier trait value (**b**) and mean dominance coefficient of deleterious gene alleles (**c**). Results are given for 4 values of recombination rate between gene and enhancer loci: $R_{EA} = 10^{-6}$ (red squares), $R_{EA} = 10^{-4}$ (purple triangles), $R_{EA} = 10^{-2}$ (green diamonds) and $R_{EA} = 0,5$ (blue discs). Three regimes appear. At low recombination, enhancer strength escalates, mean dominance coefficients remains around *0.25*, and ASE suppressor does not evolve. At intermediate recombination rate, enhancer strength escalates (though more slowly), mean dominance is over *0.25*, and ASE suppressors invade. At high recombination rate, enhancer strength does not escalate, mean dominance is over *0.25* and same ASE suppressors evolve.

On Fig. 1, we see that there can be 3 scenarios. First, when recombination rates are low ($R_{EA} = 10^{-6}$ and $R_{EA} = 10^{-4}$ on Fig. 1), we see that ER process takes place: mean enhancer strength escalates over time. This escalation does not lead to an increase of mean dominance of deleterious alleles. As a result, ASE suppressors do not evolve. In a second scenario, at intermediate recombination rates ($R_{EA} = 10^{-2}$ on Fig. 1), ER process occurs, although escalation is slower. However, this time, the recombination rate is too large for maintaining very strong linkage disequilibrium. As a result, and following what we explained previously, mean dominance of deleterious alleles increases. This causes the evolution of ASE suppressors which minimize the increased mean dominance. Finally, in a third scenario, at high rates of recombination ($R_{EA} = 0,5$ on Fig. 1), the ER process does not take place since genetic association between the gene and its regulatory region is too low to cause a significant indirect selection. However, the occurrence of enhancer polymorphism (maintained by recurrent mutations) leads to increased mean dominance, which in turn promotes the invasion of ASE suppressors.

Stochastic simulations with stabilizing selection and regulatory coevolution

We ran simulations taking into account potential costs of overexpression caused by the ER process. We used four intensities of stabilizing selection: $I = 0$ (no stabilizing selection), $I = 10^{-4}$, $I = 10^{-3}$, and $I = 10^{-2}$. To scale the intensity of stabilizing selection on expression levels to a realistic value, we can calculate the intensity of the stabilizing selection I^* that would correspond to a case where the fitness effect of producing half the optimal amount of proteins equals the mean fitness effect of deleterious alleles when heterozygotes:

$$e^{-I\left(\frac{Q^*}{2}\right)^2} = 1 - h\,s \qquad (13)$$

Throughout the simulations, parameter values are set as: $Q^* = 2$, $h = 0.25$ and $s = 0,1$. Given these values, we calculate $I^* \approx 0.025$. We consider here only cases where producing half the optimum amount of proteins is less costly than producing the optimum amount of proteins, half of which being defective. In other words, there is no fitness benefit in producing defective proteins. If such an assumption is plausible, it probably corresponds to rather low

intensity stabilizing selections, even though we lack empirical estimation of stabilizing selection intensities on expression levels and the fitness costs resulting from the production of a protein.

Concerning the model A with two enhancers co-regulating the gene, we see on Fig. 2 that coevolution depends on the balance between stabilizing selection and ER process at the more distant enhancer locus (**M**). When the modifier **M** recombines freely with the enhancer-gene pair (Fig 2.c), it does not undergo the ER process. Strength of the **M** enhancer thus tends to decrease to compensate for the enhancer strength increase due to the ER process on the **E** enhancer locus, close to the gene (R_{EA} is set at 10^{-6}) (Fig 2.d). This is not the case when $I = 0$, as there is then no need to maintain an optimal amount of protein. When the **M** enhancer is more tightly linked to the **E** enhancer and the gene (Fig 2.a), the ER process also occurs on the **M** enhancer. As a consequence, when there is no stabilizing selection ($I = 0$), **M** strength increases. It increases less than the **E** enhancer strength (Fig. 2.b), as it recombines more with the gene (intensity of ER process decreases with genetic distance between the enhancer and the gene). With stabilizing selection, **M** enhancer strength decreases less than in the free recombination scenario due to the ER process. This also reduces escalation rate for **E** enhancer. In both cases, lower stabilizing selection intensity favors faster decrease of **M** enhancer strength and faster increase of **E** enhancer strength. Individuals with non-optimal expression levels can indeed reach appreciable frequency, which allows compensatory mutations restoring optimal expression in these individuals to appear in a favourable genetic background and spread.

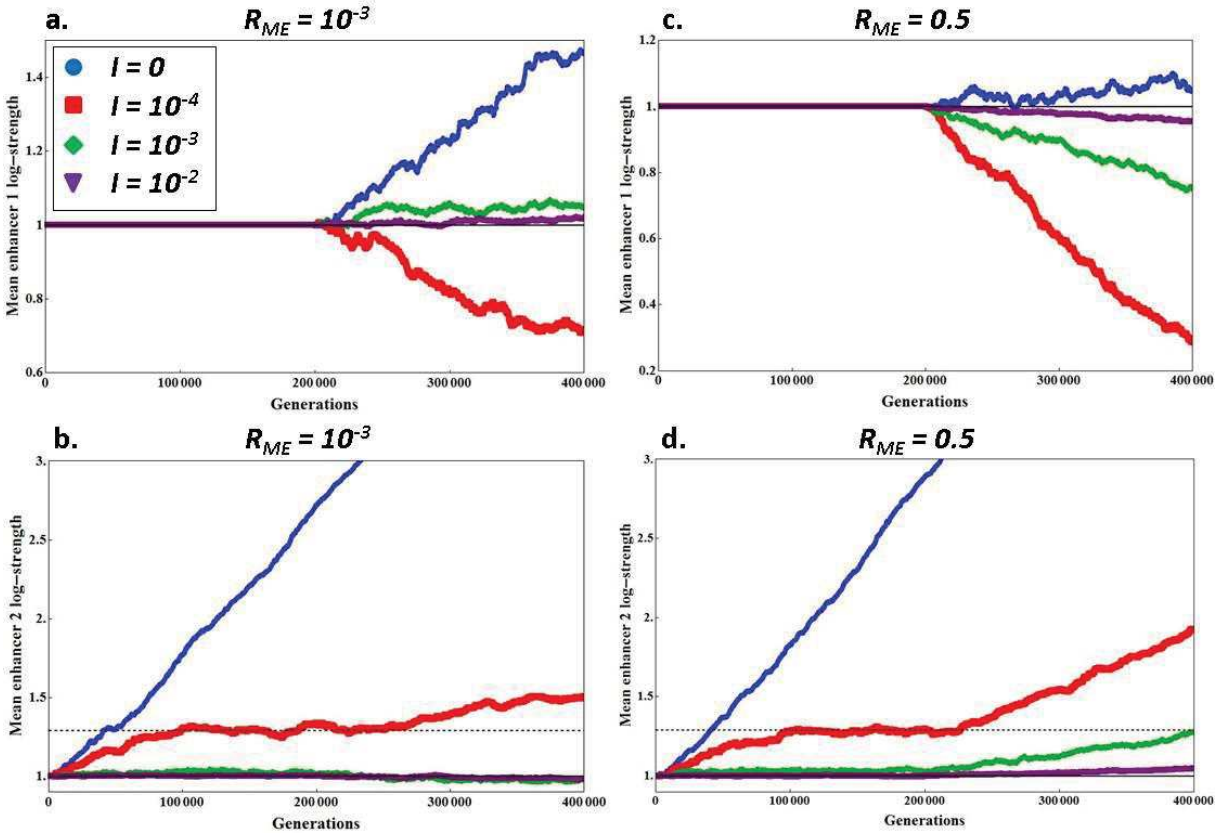


Figure 2. Simulated evolution of one gene and two enhancers (Model A) over 400 000 generations. Recombination rate between first (**M**) and second (**E**) enhancers is equal to R_{ME}: $R_{ME} = 10^{-3}$ (**a** and **b**), or $R_{ME} = 0{,}5$ (**c** and **d**). We follow mean enhancer log-strength of enhancers **M** (**a** and **c**) and **E** (**b** and **d**). Recombination rate between second enhancer and gene locus is $R_{EA} = 10^{-6}$. Total expression, calculated as first enhancer log-strength plus second enhancer log-strength on both homolog chromosomes (see Methods), undergoes stabilizing selection. Optimum value is set at 2 (1 per chromosome). Intensity of stabilizing selection is denoted I: $I = 0$ (no stabilizing selection, for comparison, blue discs), $I = 10^{-4}$ (red squares), $I = 10^{-3}$ (green diamonds) and $I = 10^{-2}$ (purple triangles). **M** enhancer only evolves after the first 200 000 generations. Until then, **E** enhancer mean log-strength is the result of an equilibrium between ER process and stabilizing selection. Dash lines on **b** and **d** show that such equilibrium for $I = 10^{-4}$ is about *1.25* (for a total expression of *2.5*). Such equilibrium for higher stabilizing selection intensity are not illustrated because too close from *1*. **M** enhancer evolution allows **E** enhancer to escape this equilibrium value.

In model B, the expression modifier is a *trans*-regulator. In this case, we observe that stabilizing selection tends to decrease the strength of the transcription factor **M** (Fig. 3.b) to maintain optimal amount of proteins while enhancer strength increases due to ER process (Fig. 3.a). Here again, coevolution is easier with low stabilizing selection intensities.

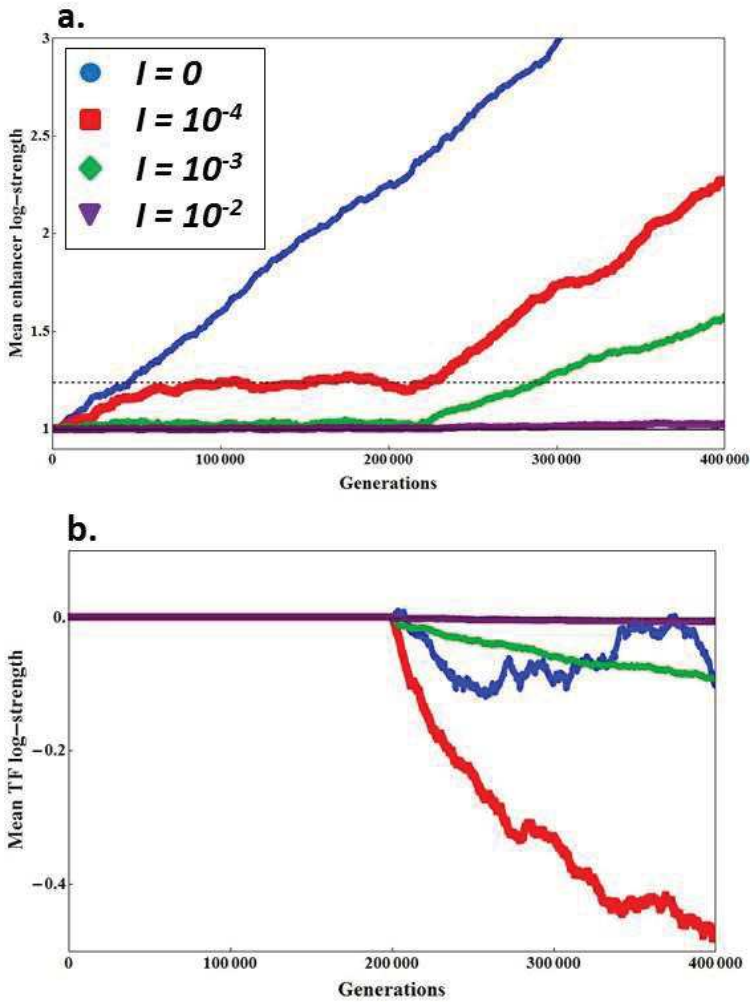


Figure 3. Simulated evolution of one gene, one enhancer and one transcription factor (Model B) over 400 000 generations. We follow mean enhancer log-strength (**a**) and mean transcription factor log-strength (**b**). Transcription factor locus is located on another chromosome ($R_{ME} = 0,5$). Enhancer locus is located close from the gene ($R_{EA} = 10^{-6}$). Total expression, calculated as enhancer log-strength times transcription factor strength on both homologous chromosomes (see Methods), undergoes stabilizing selection. Optimum value is set at 2. Intensity of stabilizing selection is denoted *I*: $I = 0$ (no stabilizing selection, for comparison, blue discs), $I = 10^{-4}$ (red squares), $I = 10^{-3}$ (green diamonds) and $I = 10^{-2}$ (purple triangles). Transcription factor locus can only evolve after the first 200 000 generations. Until then, mean enhancer log-strength is the result of an equilibrium between the ER process and stabilizing selection. Dash line on **a** shows that such equilibrium for $I = 10^{-4}$ is about *1.25* (for a total expression of *2.5*). Such equilibrium for higher stabilizing selection intensity are not illustrated because too close from *1*. Transcription factor evolution allows enhancer to escape this equilibrium value.

In model C, where a negative regulatory feedback loop can evolve, Fig. 4 shows that coevolution has a strong impact: after some time, the ER process resume at normal speed (i.e. as if there was no stabilizing selection). Indeed, the feedback loop is strongly favored, and when this regulation becomes tight enough, protein expression is maintained close to its optimum value, irrespectively of mean enhancer strength. In such cases, impact of stabilizing selection on the enhancer locus is very small, and consequently the rate of the ER process is unaltered. Importantly, selection of a negative feedback loop is stronger when stabilizing selection is more intense. Here, any mutations that decreases the modifier trait value (increase the tightness of the feedback loop) is beneficial, which was not the case in the two previous models. In models A and B, a modifier mutation that was over-compensating was deleterious; explaining why coevolution was slower with stronger stabilizing selection.

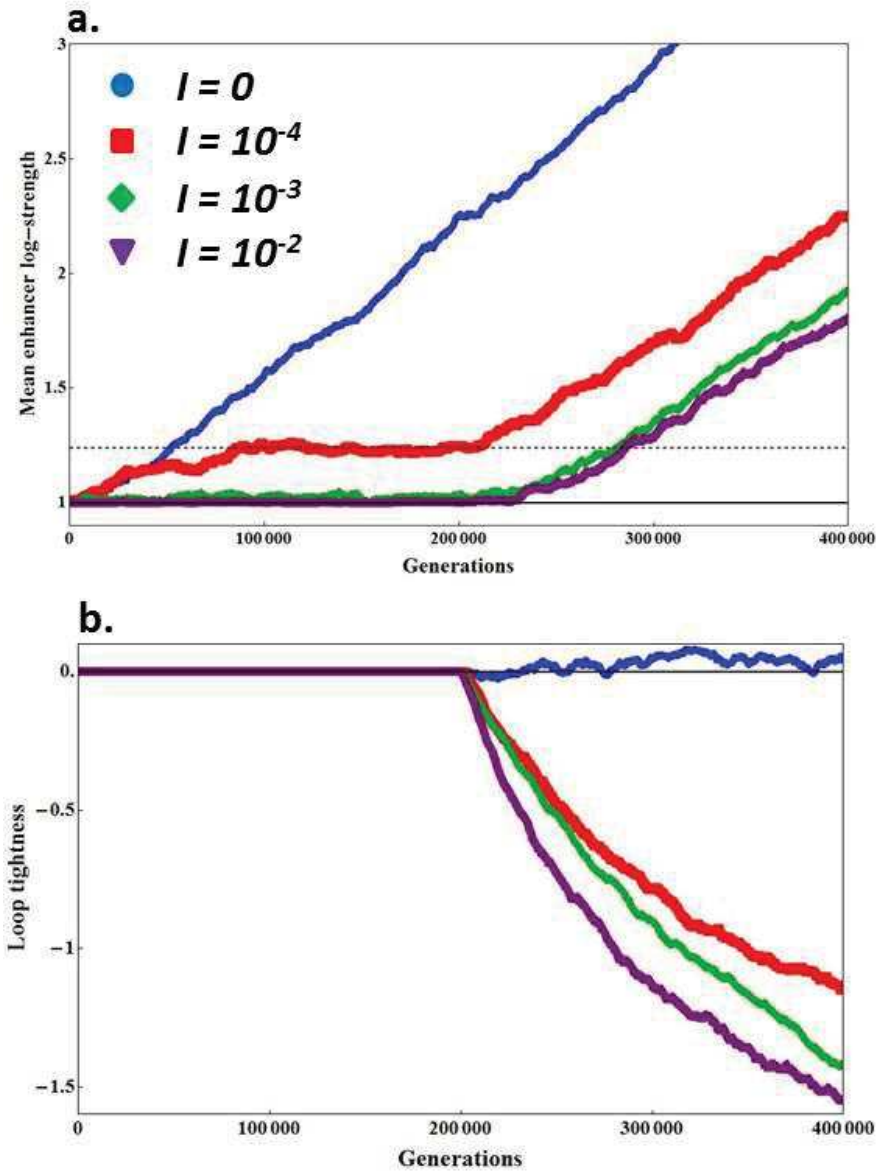


Figure 4. Simulated evolution of one gene, one enhancer and one *trans*-regulatory sequence implementing a negative feedback regulatory loop (Model C), over 400 000 generations. We follow mean enhancer log-strength (**a**), and mean loop tightness (loops are tighter for negative modifier trait values) (**b**). Loop sequence is located on another chromosome (R_{ME} = *0,5*). Enhancer locus is located close from the gene ($R_{EA} = 10^{-6}$). Total expression, calculated as described in Methods, undergoes stabilizing selection. Optimum value is set at 2. Intensity of stabilizing selection is denoted *I*: $I = 0$ (no stabilizing selection, for comparison, blue discs), $I = 10^{-4}$ (red squares), $I = 10^{-3}$ (green diamonds) and $I = 10^{-2}$ (purple triangles). Regulatory loop locus can only evolve after the first 200 000 generations. Until then, mean enhancer log-strength is the result of an equilibrium between the ER process and stabilizing selection. Dash line on **a** shows that such equilibrium for $I = 10^{-4}$ is about *1.25* (for a total expression of *2.5*). Such equilibrium for higher stabilizing selection intensity are not illustrated because too close from *1*. Negative feedback loop evolution allows enhancers to escape this equilibrium value, and resume normal strength escalation.

Discussion

Coevolution in the regulome

The ER process concerns only enhancers relatively close to their genes. In FYON *et al.* (2015), we showed that there is a genetic distance window around genes where the ER process occurs. The limit of this window depends on several parameters, most importantly the intensity of purifying selection at the gene locus and the size of enhancer mutations. We show here that the consequences of the ER process extend far beyond this window around the gene, and impact *cis*-regulatory sequences as well as *trans*-regulatory sequences.

There may be different costs associated to the ER process. The most obvious one is the cost of over-expressing proteins due to ever-stronger enhancers. Our results show that the simplest way to compensate for over-expression is to evolve negative feedback regulatory loops. These loops allow for fine tuning of expression levels, through the negative retro-action of a molecule produced or activated downstream of gene expression. With such regulation, expression levels are very stable against genetic or stochastic variability. ER process can then take place as it only relies on relative expression differences (differences of expression between homologous copies, on which negative feedback regulation has no impact), but expression levels remain unchanged despite stronger enhancers. Note that we did not consider possible costs arising from the evolution of a regulatory loop, such that we may expect these loops to mostly concern genes under strong stabilizing selection. Without explicit measures of these costs, it may be difficult to predict the extent to which regulatory loops should evolve. Also, some genes may not be exposed to stabilizing selection on expression: positive loops can be selected for to cause abrupt on/off switches (FERRELL JR 2002), and some situations have been argued to favor expression noise (SEGER and BROCKMAN 1987; BLAKE *et al.* 2003; VINEY and REECE 2013). We did not consider these cases here.

Our results show that the ER process impacts all enhancers (see also Annexe). For enhancers tightly linked to the gene, indirect selection for increased enhancer strength is stronger than stabilizing selection and enhancer strength increases. For more distant enhancers, indirect selection becomes weaker, and is overcome by stabilizing selection. In these regions, enhancer strength decreases, which allows for the compensation of enhancer strength escalation closer to the genes. Because of this coevolution, we expect a gradient of enhancer

strength with distance from the gene. This is also illustrated in Annexe. However, this pattern depends on the intensity of stabilizing selection. Indeed, if stabilizing selection is so intense that nearly no temporary non-optimal expression is allowed, coevolution between cis-regulatory sequences is not possible, and the rate of ER process becomes negligible. In the other hand, with such a strong stabilizing selection, we expect negative feedback regulatory loop to evolve, which would then allow the ER process to proceed at a high rate.

Transcription factors can also coevolve with enhancers to compensate for overexpression, if the stabilizing selection is not too intense. Our results show that their possibilities to coevolve to compensate for over-expression are similar to those of distant enhancers. However, TF evolution may be seriously limited by their pleiotropic effects. TFs are indeed usually involved in expression regulation of many genes (WRAY *et al.* 2003). If decreasing the activity of a TF is beneficial for fine-tuning expression of one gene, it may be deleterious for another gene whose regulation also depends on that TF. Overall, if the ER process could impact *trans*-regulatory elements, it is likely to be of lesser importance than coevolution of *cis*-regulatory sequences.

The other cost of the ER process concerns average dominance. In much of the recombination window where the ER process can occur, it induces a lower average dominance of deleterious gene alleles: because deleterious alleles tend to be associated with weaker enhancers, they are better hidden. However, around the recombination limit of this window, average dominance of deleterious mutations increases. This does not prevent stronger enhancers to spread, as they are still associated with good genetic background, only not enough for mean dominance to increase. This is costly, and as a consequence ASE suppressors can spread to reduce this increase of mean dominance. ASE suppressors equalize expression between homologous gene copies, and reduce allele-specific expression. This could correspond to structural genomic changes, as the evolution of homolog pairing and transvection in *Diptera* species (MELLERT and TRUMAN 2012) or homolog repulsion in mammals (HERIDE *et al.* 2010). When this ASE suppressor has a global structural effect, like in these examples, its selective advantage sum up on many loci and could be quite large. Interestingly, we also show that such modifiers should also be selected for even if the ER process does not operate, or if enhancers are not within the recombination window where it

can occur. Indeed, even when stronger enhancers do not spread, enhancer mutational polymorphism tends to increase mean dominance.

Some papers have argued that neutral evolution has a significant impact on regulatory networks evolution (KHAITOVICH *et al.* 2004; YANAI *et al.* 2004; WHITEHEAD and CRAWFORD 2006). This mode of evolution is favored by the availability of many, functionally equivalent, regulatory networks. Evolution between such networks should thus be neutral or near-neutral, if we assume that intermediary networks do not disrupt expression levels too much. Our results shed a new light on this process: evolution may actually not be neutral between similar networks. In a set of similar networks, our three-locus model shows that indirect selection for stronger enhancers should favor the networks with the strongest enhancers at locations close from the gene. Our architecture model (see Annexe) shows that competition for expression should favor networks with more enhancer sequences, and networks with strong enhancers close from the gene and silencing enhancers further away. This advantage of stronger enhancers close from the gene should stimulate evolution between similar regulatory networks, if it exceeds the cost of intermediary networks (that lead to sub-optimal expression levels). Indirect selection on enhancers close from the gene, and then on other regulatory sequences, may strongly influence, on the long term, the structure and complexity of eukaryotic networks and explain why regulatory networks could change quickly without bearing adaptive novelties (TAUTZ 2000).

A process that shares many similarities with genetic conflicts

Genetics conflicts are defined as the divergence of evolutionary benefits of different genetic sequences in one genome (HURST *et al.* 1996; BURT and TRIVERS 2009). One sequence is called the 'selfish element'. It spreads gaining a better transmission to the next generation, without increasing the fitness of the host individual, or even decreasing it. They increase their own transmission rates and decrease the transmission rates of other sequences in the host. A distinctive feature of selfish genetic elements is that their presence triggers the evolution of 'suppressors' that limit or prevent their spread, and restore individual's fitness.

The ER process shares many similarities with genetic conflicts. Indeed, stronger enhancers competing for expression can be seen as selfish elements. They spread by getting associated with good genetic background, which may not be beneficial for the whole organism. It may

result in the two types of costs for organisms mentioned above: over-expression of proteins and increased average dominance of deleterious gene alleles.

To respond to over-expression costs, we see that organisms evolve co-regulators to maintain optimal expression levels. They evolve in response to the spread of stronger enhancers. However, they do not stop stronger enhancers from invading. On the contrary, they allow for stronger enhancers to keep invading despite the stabilizing selection. Thus, they cannot be referred to as 'suppressors'. The term 'compensators' suits them better. Here, 'suppressors' are not the only way to reduce the individual fitness costs associated to the presence of selfish elements. There is no conflict here, but conciliation. Rather than preventing stronger enhancers from invading, the whole regulome evolves to continually compensate induced costs.

To respond to dominance costs (which do not systematically occur with ER process), ASE suppressors can evolve. They decrease mean dominance by reducing allele-specific expression. This tends to progressively stop the ER process, which is based on allele-specific expression. However, ASE suppressors can evolve for other reasons than suppressing the ER process. Enhancer mutational polymorphism alone, without escalation in enhancer strength, can cause an increased average dominance and favor the spread of ASE suppressors. There is not a one to one correspondence between selfish enhancers and suppressors to control them, as in a basic situation of genetic conflict. ER process may not necessarily favor the evolution of ASE suppressors and ASE suppressors may evolve for other reasons than controlling the spread of selfish enhancers.

If we except the incorporation of viral or transposable element in regulatory regions (RICE 2013), the evolution of gene expression has not been understood to date as resulting from genetic conflicts (BURT and TRIVERS 2009). We showed here that original kind of genetic conflicts can be involved in the evolution of gene expression. These findings reinforce the view that "nothing in genetics makes sense except in light of genomic conflicts" (RICE 2013).

References

BLAKE W. J., KAERN M., CANTOR C. R., COLLINS J. J., 2003 Noise in eukaryotic gene expression. Nature **422**: 633–637.

BURT A., TRIVERS R., 2009 *Genes in conflict: the biology of selfish genetic elements*. Harvard University Press.

FAY J. C., WITTKOPP P. J., 2008 Evaluating the role of natural selection in the evolution of gene regulation. Heredity (Edinb). **100**: 191–9.

FERRELL JR J. E., 2002 Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability. Curr. Opin. Cell Biol. **14**: 140–148.

FYON F., CAILLEAU A., LENORMAND T., 2015 Enhancer Runaway and the Evolution of Diploid Gene Expression. PLoS Genet **11**: e1005665.

HERIDE C., RICOUL M., KIÊU K., HASE J. VON, GUILLEMOT V., CREMER C., DUBRANA K., SABATIER L., 2010 Distance between homologous chromosomes results from chromosome positioning constraints. J. Cell Sci. **123**: 4063–4075.

HURST L. D., ATLAN A., BENGTSSON B. O., 1996 Genetic conflicts. Quaterly Rev. Biol. **71**: 317–364.

KHAITOVICH P., WEISS G., LACHMANN M., HELLMANN I., ENARD W., MUETZEL B., WIRKNER U., ANSORGE W., PÄÄBO S., 2004 A Neutral Model of Transcriptome Evolution. PLoS Biol. **2**: e132.

KUO D., LICON K., BANDYOPADHYAY S., CHUANG R., LUO C., CATALANA J., RAVASI T., TAN K., IDEKER T., 2010 Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Res.: 1–7.

LUDWIG M. Z., BERGMAN C., PATEL N. H., KREITMAN M., 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**: 564–7.

MANNA F., MARTIN G., LENORMAND T., 2011 Fitness Landscapes: An Alternative Theory for the Dominance of Mutation. Genetics **189**: 923–937.

MELLERT D. J., TRUMAN J. W., 2012 Transvection Is Common Throughout the Drosophila Genome. Genetics **191**: 1129–1141.

RICE W. R., 2013 Nothing in Genetics Makes Sense Except in Light of Genomic Conflict.

Annu. Rev. Ecol. Evol. Syst. **44**: 217–237.

SEGER J., BROCKMAN H. J., 1987 What is bet-hedging ? Oxford Surv. Evol. Biol. **4**: 182 – 211.

TAUTZ D., 2000 Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. **10**: 575–579.

VINEY M., REECE S. E., 2013 Adaptive noise. Proc. R. Soc. B Biol. Sci. **280**: 20131104.

WEIRAUCH M. T., HUGHES T. R., 2010 Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. **26**: 66–74.

WHITEHEAD A., CRAWFORD D. L., 2006 Neutral and adaptive variation in gene expression. Proc. Natl. Acad. Sci. **103**: 5425–5430.

WRAY G. A., HAHN M. W., ABOUHEIF E., BALHOFF J. P., PIZER M., ROCKMAN M. V, ROMANO L. A., 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–419.

YANAI I., GRAUR D., OPHIR R., 2004 Incongruent Expression Profiles between Human and Mouse Orthologous Genes Suggest Widespread Neutral Evolution of Transcription Control. Omi. A J. Integr. Biol. **8**: 15–24.

L'*ER process* semble pouvoir être à l'origine de nombreuses modifications dans les réseaux de régulation de l'expression des gènes. Il tend à augmenter la complexité du réseau, et à favoriser une divergence rapide des séquences régulatrices. Il agit comme moteur d'un remplacement des séquences régulatrices sans changement phénotypique : le réseau évolue pour compenser l'augmentation de la force des promoteurs proches des gènes et maintenir des niveaux d'expression optimaux. Bien que la pression de sélection indirecte sur les séquences régulatrices à la base de l'*ER process* soit faible (de l'ordre du taux de mutation), elle peut être significative en grande population. Elle agit de plus probablement depuis l'apparition des eucaryotes (l'origine des eucaryotes est supérieure au milliard d'année). Il est donc tout à fait envisageable, sur le long terme, que l'*ER process* ait participé à la formation des réseaux de régulation tels qu'on les voit maintenant.

Quatre paramètres semblent particulièrement influencer l'*ER process* : l'intensité de la sélection purifiante sur le gène, le taux de mutation du gène, le taux de recombinaison entre le gène et le promoteur et l'intensité de la sélection stabilisante sur les niveaux d'expression. Un cinquième paramètre, que nous n'avons pas encore étudié, est également important : l'hétérozygotie. En effet, la sélection indirecte sur les réseaux de régulation ayant lieu chez les hétérozygotes, on s'attendrait à ce que des niveaux d'hétérozygotie plus forts soient associés à des *ER process* également plus forts. Toutefois, le taux de recombinaison efficace dépend aussi de l'hétérozygotie : plus il y a d'hétérozygotes, plus il y a d'événements de recombinaison menant à un échange d'allèles, ce qui tend à ralentir l'*ER process*. La relation entre vitesse d'escalade de la force des promoteurs et hétérozygotie n'est donc pas triviale.

Afin d'étudier cette question, nous avons implémenté dans nos modèles différents modes de reproduction : reproduction croisée, autogamie, parthénogénèse (différents sortes) et reproduction mitotique. Nous nous sommes rendu compte que la vitesse de l'*ER process* est positivement corrélée à l'hétérozygotie de la population. Cette corrélation est la même quel que soit le mode de reproduction, hormis dans les lignées clonales. Dans ces cas, un nouveau processus émerge : le processus de divergence des promoteurs (*ED process*).

Du fait de l'isolement génétique des chromosomes homologues dans ces lignées clonales, les promoteurs forts ne peuvent envahir complètement la population, et sont restreints aux chromosomes qui les ont vus apparaître. Les associations génétiques entre promoteurs forts

et allèles viables s'accumulent alors sur un des deux homologues, tandis que les associations entre promoteurs faibles et allèles délétères se concentrent sur l'autre homologue. Ainsi, pour chaque gène, une copie est purgée des allèles délétères et est fortement exprimée, tandis que l'autre copie accumule des allèles délétères et est peu exprimée. Cette divergence mène progressivement à l'haploïdisation de l'expression des lignées clonales.

La transition entre *ER* et *ED process* n'est pas une frontière bien délimitée, mais un régime intermédiaire. Au fur et à mesure que l'isolement génétique des chromosomes homologues diminue, les copies homologues divergent de moins en moins, et les promoteurs augmentent en force de plus en plus.

Article submitted in Genetics

Chapitre III

Enhancer Runaway and Enhancer Divergence in Asexuals

Authors : Frédéric Fyon, Thomas Lenormand

UMR 5175 CEFE, CNRS - Université Montpellier - Université P. Valéry - EPHE, 1919 route de Mende 34293 Montpellier Cedex 5, France

Abstract

With the advent of new sequencing technologies, the evolution of gene expression is becoming a subject of intensive genomic research, with sparkling debates upon the role played by these kinds of changes in adaptive evolution and speciation. In this paper we model expression evolution in species differing by their reproductive systems. We consider different rate of sexual versus asexual reproduction and the different type of parthenogenesis (apomixis and the various modes of automixis). We show that competition for expression leads to two selective processes on *cis*-regulatory regions that act independently to organism-level adaptation. Coevolution within regulatory networks allows these processes to occur without strongly modifying expression levels. First, *cis*-regulatory regions such as enhancers evolve in a runaway fashion because they automatically become associated to chromosomes purged from deleterious mutations ("Enhancer Runaway process"). Second, in almost pure asexual species, homologous *cis*-regulatory regions tend to diverge, which leads to haploidization of expression, when they are sufficiently isolated from one another ("Enhancer Divergence process"). We show how these two processes co-occur and vary depending on the level of sex and heterozygosity. This study offers thus a baseline to understand patterns of expression evolution across the diversity of eukaryotic species.

Introduction

The regulation of gene expression controls many aspects of phenotypes. Regulatory elements and their network of interactions and signaling determine where and when proteins are expressed; they regulate in which quantity they are produced; they can even influence their final folding. Evolution of the regulome undoubtedly contributes to adaptation (CARROLL 2005): many studies have observed changes in regulatory elements or networks that correlate with adaptive novelties, in bacteria (COOPER 2003), yeasts (FEREA *et al.* 1999), insects (RAYMOND *et al.* 1998), fishes (SHAPIRO *et al.* 2004), birds (ABZHANOV *et al.* 2004) to cite a few. Stabilizing selection on levels of expression have also been widely documented (WHITEHEAD and CRAWFORD 2006b LUDWIG *et al.* 2000; DENVER *et al.* 2005; GILAD *et al.* 2006; FAY and WITTKOPP 2008).

Gene expression regulators belong to two categories: *cis*-acting DNA sequences (enhancers, repressors, insulators, 5'UTRs, and core promoters) and *trans*-acting RNA or protein intermediates (transcription factors and cofactors). *Cis*-acting sequences regulate gene copies located on the same chromosome (and not on the homolog chromosome). Transcription factors (TFs) regulate in *trans*, i.e. genes on both chromosomes (for a review of gene expression regulatory architecture, see WRAY *et al.* 2003).

The central role of *cis*-regulatory changes in adaptation has been strongly advocated (WITTKOPP *et al.* 2004; CARROLL 2005; WRAY 2007). In this view, *cis*-regulatory sequences disproportionately contribute to adaptation because they are thought to have a co-dominant effect on transcript abundance, which make them more visible to selection than coding sequence mutations (WRAY 2007) ; and because they are thought to be organized in distinct modules, each controlling a different aspect of expression profiles, which should limit mutation pleiotropy (CARROLL 2008; WITTKOPP and KALAY 2012). This view is controversial and may be premature given the lack of strong theoretical and empirical validation (HOEKSTRA and COYNE 2007; LYNCH and WAGNER 2008). Alternatively, many regulatory changes have been argued to evolve (quasi) neutrally (OLEKSIAK *et al.* 2002; KHAITOVICH *et al.* 2004; YANAI *et al.* 2004). This is because some regulatory mutations may have no effects on expression profiles and because different networks can result in similar expression profiles, leading to similar phenotypes (WEIRAUCH and HUGHES 2010). The fitness landscape associated to regulatory

networks is likely to present fitness ridges rather than a single well defined peak as considered in stabilizing selection scenarios: different networks may lead to the same optimal phenotype. Most regulation indeed relies on the proper matching between several regulatory components whose precise identity or quantity is somehow arbitrary as long as they match (as in signal/receiver, or key/lock situations), allowing for considerable "evolutionary freedom" (LENORMAND *et al.* 2015). Evolutionary divergence along those ridges can occur by the spread of slightly deleterious mutations and the occurrence of compensatory mutations restoring optimal expression profiles of targeted genes (TAUTZ 2000; KUO *et al.* 2010; WEIRAUCH and HUGHES 2010; COOLON *et al.* 2014). The structural features of regulatory networks (complexity, connectivity, redundancy, …) may also be largely shaped by neutral processes (FORCE *et al.* 2005; LYNCH 2007).

The debates over the role of regulatory changes in Evolution have strongly focused on their role for organismal adaptation. Yet, self-promoting regulatory sequences may evolve in a runaway process even without optimizing individual level traits, which considerably complicate the evaluation of their role in adaptation as usually envisioned. Furthermore, this process generates selection pressures that are not considered in neutral models. We described this process recently, by explicitly modelling mutations in regulatory (*cis*- or *trans*-) or genic regions in diploids (FYON *et al.* 2015). In brief, this runaway occurs when *cis*-regulators 'compete' for expression. We used enhancers as example of *cis*-regulators, and showed that stronger enhancers (enhancers that activate more transcription) tend to get preferentially associated with good genetic background. Indeed, stronger enhancers express a larger share of proteins than their weaker homologs. As a result, their associated gene copy (in *cis*) tends to contribute more to final phenotype, is more exposed to selection and thus better purged from deleterious mutations. This association between stronger enhancers and good genes allows for the stronger enhancers to invade. Overall, enhancers' strength escalates, leading to an endless 'Enhancer Runaway' (ER) process.

This process should occur very generally for most genes across the genomes of most eukaryotes exposed to selection during their diploid phase. It's expected to happen however only for enhancers close enough from their target genes (so that the positive genetic association between the stronger enhancer and the purged gene is not broken down too often by recombination). It is important to note that the ER process would not necessarily

lead to increased expression levels, which would be deleterious. Due to the occurrence of multiple regulatory networks ensuring the same regulation, coevolution between regulators could ensure optimal expression levels despite ever-stronger enhancers (Fyon & Lenormand, in prep). For instance, the spread of stronger enhancers could be compensated by the concomitant spread of a 'weaker' associated transcription factors. Because of the ER process, evolution through those multiple networks is expected to be not neutral, but biased for networks with strong enhancers close to the genes. Hence, the evolution of regulatory regions may not be simply driven by optimization of gene expression levels, or quasi-neutral dynamics, but also by the endless spread of self-promoting *cis*-regulators. ER process may also have contributed to shape many features of expression control and regulatory networks (see discussion in FYON *et al.* 2015). For example, some of the complexity of regulatory architecture might stem from the accumulation of *cis*-regulatory elements driven by the ER process.

The rate of the ER process depends on the proportion of heterozygotes in the population. As a consequence, we might expect reproductive systems reducing this proportion to slow it down. This is indeed the case for self-fertilization (FYON *et al.* 2015). Furthermore, enhancer runaway may occur differently in asexuals where recombination is reduced and where stronger enhancer may not be able to escape from their lineage of origin. To obtain a more comprehensive view of the dependence of the ER process on the reproductive mode, we investigate in this paper how the ER process varies in a wide variety of such systems, considering in particular the various forms of automixis and apomixis, found in parthenogenetic species (see ASHER 1970; SUOMOLAINEN and LOKI 1987; NOUGUÉ *et al.* 2015 ; and SCHÖN *et al.* 2009 for technical details about these reproductive systems). This may serve as a basis for the comparative analysis of the evolution of regulatory regions across sister species differing by their mode of reproduction and shed light on the evolution of regulatory regions in asexuals.

Methods

The model builds on the individual-based stochastic model used in FYON *et al.* 2015. Individuals' genomes are represented by two loci: one gene and its associated *cis*-regulatory region (e.g. an enhancer locus). *Cis*-regulatory sequences include many elements, like enhancers, core promoters or 5'UTRs. Results derived from the model are valid for any type of *cis*-regulatory sequence, provided that it can mutate and alter expression levels. We often use the term enhancers simply because they represent typical *cis*-regulatory sequences that can be found at different recombination distances from their target gene (WRAY 2007). The process we study here highly depends on linkage disequilibrium between the gene and the *cis*-regulator, so it will act differently at different genetic distances, for different enhancers.

Individuals go through a simple life-cycle: diploid selection, meiosis with recombination, mutations and syngamy.

Selection

In this model, fitness depends on the presence of deleterious alleles on the gene and on their relative expression levels. However, we assume that total (absolute) expression levels do not influence fitness. This is similar to assuming that the gene is embedded into a negative feedback loop, which ensures constant optimal absolute expression levels. This assumption allows us to focus on the effects of expression asymmetry between homologs without the confounding effects of selection on absolute expression levels. Studying relative expression only requires focusing on cis-regulators. Studying the interplay between relative (between alleles) and absolute expression levels is beyond the scope of this paper, as it would require considering more loci (other cis-regulators, trans-acting regulators, loop regulators), as several regulators are necessarily involved in total expression levels. This topic is tackled in another publication in preparation (Fyon & Lenormand, in prep).

The fitness of an individual *i* is calculated as:

$$W_i = w_{i,1} + h_i\left(w_{i,2} - w_{i,1}\right), \qquad (1)$$

where $w_{i,1}$ is the fitness of the fittest gene allele of individual *i*, $w_{i,2}$ the fitness of the other gene allele of individual *i*, and h_i the dominance coefficient of the least fit allele in individual *i*. Note that when $h_i = 0.5$ then $W_i = (w_{i,1} + w_{i,2})/2$. Without loss in generality, we suppose that dominance coefficients depend on the relative expression level of the least fit allele, which depends on the relative strengths of homologous enhancers. We express dominance coefficients using the following relationship, which ensures that dominance of an allele increases with its relative expression and ensure that dominance coefficient is always equal to the parameter *h* for equally expressed alleles:

$$h_i = \left(\frac{e_{i,1}}{e_{i,1}+e_{i,1}}\right)^{-\frac{Log(h)}{Log(2)}}, \quad (2)$$

where $e_{i,1}$ and $e_{i,2}$ are the strengths of enhancers associated with gene alleles of fitness $w_{i,1}$ and $w_{i,2}$ respectively, and *h* being the dominance coefficient in individuals that are homozygotes at the enhancer locus (and where both gene alleles are thus equally expressed). In the following, we set *h = 0.25*, which corresponds to the empirical consensus for the dominance of mildly deleterious mutations (MANNA *et al.* 2012). With such partial recessivity, dominance of an allele necessarily increases more than linearly with its proportion of expression.

To model selection and reproduction, we sample with replacement individuals in the population, accepting them with a probability equal to their fitness until we obtain two parents. One (possibly recombined) chromosome (one gamete) in each parent is then sampled to form a new offspring, and this procedure is repeated until we obtain N_{pop} offspring. In each parent a recombination event between the gene and the enhancer can occur with a probability equal to the recombination rate per individual per generation R_{EA}.

Mutations

Mutations occur in gametes. The gene locus undergoes recurrent recessive deleterious mutations at a rate u_A. This is implemented by drawing the total number of mutations

occurring in the population with a Poisson distribution with a mean equals to $2N_{pop}u_A$. The fitness effect of every mutation is drawn from a negative exponential distribution of mean *s*. The enhancer locus also undergoes mutations at a rate u_E. Similarly, the number of enhancer mutations is drawn from a Poisson distribution with a mean equals to $2N_{pop}u_E$. As it is our focus, we assume that these mutations only affect expression levels of alleles of the gene. In particular, we do not consider that they have other effects (e.g. on expression timing or localization). Such mutations are expected to occur since c*is*-regulatory changes are thought to exhibit low pleiotropy (CARROLL 2005; WRAY 2007). Also, different regulatory networks can lead to similar phenotypes (WEIRAUCH and HUGHES 2010), indicating that pleiotropic effects of mutations can be easily compensated. Mutations change additively the log of enhancer strength, which ensures that the mutational effect size remains constant irrespectively of arbitrarily chosen absolute values of enhancer strength (see FYON *et al.* 2015 for details). Mutation effects are drawn from a Normal distribution of mean *0* and standard variation σ_E. These assumptions on enhancer mutations allow us to observe a linear increase of enhancer strength under ER process (FYON *et al.* 2015). Other mutational regimes can be considered, and are expected to lead to quantitative but not qualitative differences. ER process only needs mutations altering relative expression levels of homologous gene copies to occur.

Reproductive modes

In FYON *et al.* 2015, we showed that self-fertilization reduced the rate of the ER process. Here, we provide a much more comprehensive study of the impact of the reproductive modes on the ER process and we consider the case of complete absence of recombination, which leads to qualitatively very different outcomes. Different modifications of the simple life cycle described above were considered to implement different mating systems. These mating systems are presented in Table 1 and illustrated on Fig 1. They include self-fertilization, various forms of automixis and apomixis. For self-fertilization, the two chromosomes transmitted to the offspring are chosen from two independent meiosis events from a single parent individual. With automixis, the two chromosomes forming a new offspring are sampled *without replacement* among the four meiotic products of a single meiosis. The precise sampling scheme depends however on the mode of automixis. With Central Fusion, the two chromosomes sampled come from meiotic products that are split at meiosis I. With Terminal Fusion, they come from meiotic products that are split at meiosis II.

In Random Fusion, they come at random from the four meiotic products. In Pre-Meiotic Doubling, the DNA is replicated twice (instead of once in normal meiosis) before meiosis and one of the diploid products of the meiosis gives rise to a new individual. In Post-Meiotic Doubling, the genome is doubled in the meiotic products of a normal meiosis, and one of them will develop into a new individual. Finally, in apomictic reproduction, no meiosis occurs, and individuals are formed from a mitotic division in the parent (we assume no mitotic recombination). These modified life cycles occur with a given probability to each individual at every generation to model mixed mating systems.

N°	Reproductive mode	# meiosis	Details	% of heterozygosity retention
1	Apomixis	0	Mitosis	100%
2	Pre-meiotic doubling	1	Duplication of chromosomes before meiosis, identical chromosomes resulting from duplication pair during meosis I	100%
3	Post-meiotic doubling	1	Endomitosis of the meiotic product	0%
4	Central fusion	1	Central fusion in ordered tetrads or suppression of meiosis I	From 100% at centromere to 66% at large genetic distances (in Morgan) from centromere
5	Terminal fusion	1	Terminal fusion in ordered tetrads or suppression of meiosis II	From 0% at centromere to 66% at large genetic distances (in Morgan) from centromere
6	Random fusion	1	Random fusion in tetrads = mixed fusion with terminal and central fusion in proportion 2/3 and 1/3, respectively	Intermediate between central and terminal fusion
7	Self-fertilization	2	Syngamy of a male and a female gamete from two independent meioses within the same individual	50%

Table 1. Description of the different breading systems that allow a single individual to reproduce without any mate. Apomixis (1) does not involve meoisis. Automixis (2-6) involves the production of an offspring by fusion of the two products from a single meiosis (unlike self-fertilization where offspring are produced by fusion of two products from two independent meioses, in the male and the female gametes, respectively). Central and terminal fusions are usually distinguished. Central (resp. terminal) fusion corresponds to the fusion of meiotic products derived from the first (resp. second) meiotic division. Central fusion retains heterozygosity at centromere positions while terminal fusion leads to the loss of heterozygosity at centromere positions. In both cases heterozygosity is reduced by one third at positions far away from the centromeres. Thus, automixis through central fusion combined with very low recombination rates leaves a genetic signature very similar to that of apomixis (with maintenance of high level of heterozygosity). In contrast, central, terminal and mixed fusions combined with very high recombination rates leaves a genetic signature very similar to self-fertilization (with nearly complete loss of heterozygosity). This table is adapted from (NOUGUÉ *et al.* 2015).

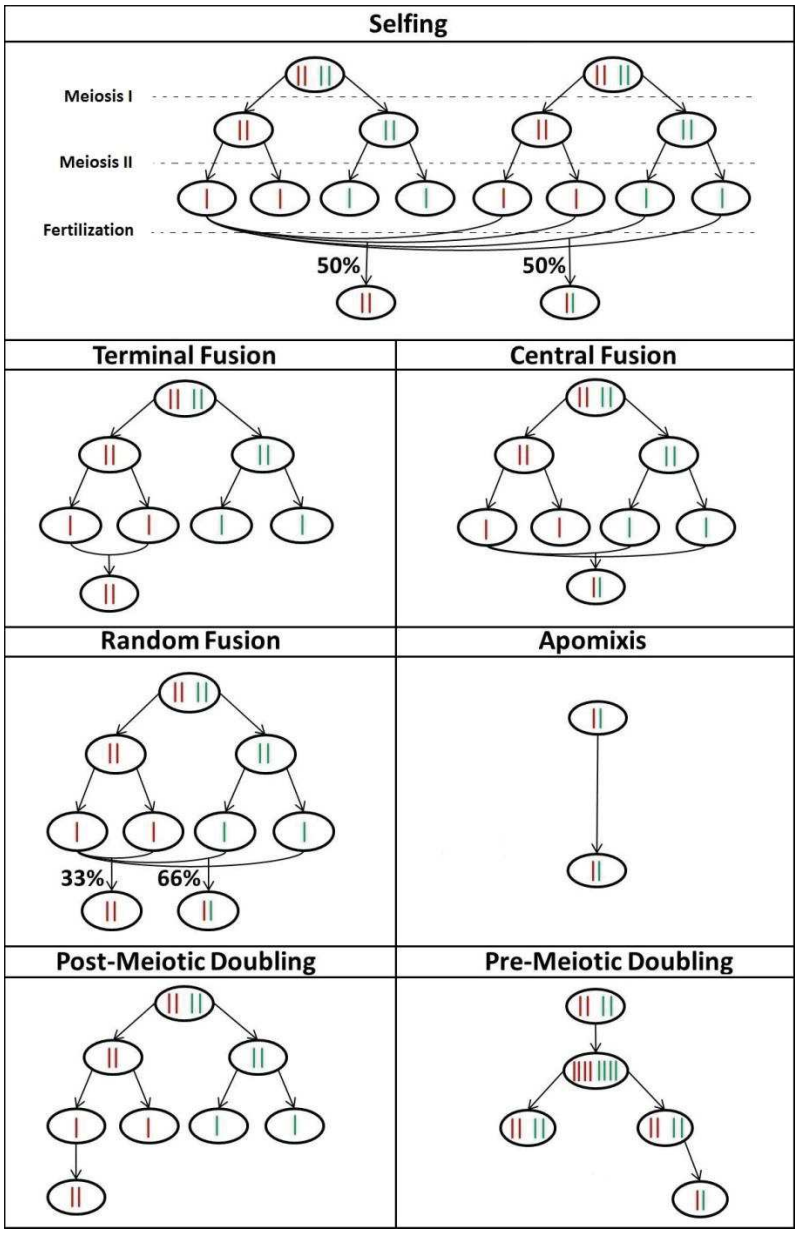


Fig. 1. Schematic illustration of the various reproductive systems implemented in our models. In Selfing, gametes from two independent meiosis fuse to form a new individual. In Automictic reproduction, gametes from the same meiosis fuse to form a new individual. Those gametes may come from the same Meiosis II division (Central Fusion), or from different Meiosis II division (Terminal Fusion). They may also randomly come from the same or different Meiosis II divisions (Random Fusion). In automictic Pre-Meiotic Doubling, DNA content is doubled before Meiosis I, such that products of meiosis are directly diploid individuals. In automictic Post-Meiotic Doubling, DNA content is doubled in gametes to give birth to a diploid individual. Finally, apomixis is a mode of reproduction where new individuals are produced mitotically, without any meiosis division.

Parameter values

Our objective is to study how cis-regulatory sequences evolve with various reproductive modes and in absence of recombination. To study this variation, we considered a typical situation where the enhancer runaway is well characterized and relatively fast (FYON *et al.* 2015): (1) high rates of mutations ($u_A = u_E = 10^{-4}$) that allow for significant polymorphism; (2) high intensity of purifying selection ($s = 0.1$) on the gene and low recombination rates ($R_{EA} = 10^{-6}$ unless under apomixis where $R_{EA} = 0$), to allow for substantial indirect selection on *cis*-regulatory sequences; (3) relatively large population size, to reduce stochastic effects of genetic drift ($N_{pop} = 10\,000$). We considered the different modes of reproduction, as described above, and for each of them, we varied the rate of non-random mating. For the special case of automixis through central fusion, we first assumed no recombination between the enhancer-gene pair and the centromere. For reasons explained below, we then relaxed this assumption and introduced some recombination with the centromere at a rate R_c. For a given set of parameter values, simulations were repeated 200 times, for 10^5-10^6 generations depending on the case.

Testing reproductive mode variation implementation in the model

To check that new mating systems have been correctly implemented in the models, we calculated expected inbreeding coefficients for populations performing given rates of selfing, automixis, or apomixis, and measured actual inbreeding coefficients in simulated populations evolving neutrally (no selection), and performing same rates of selfing, automixis and apomixis. To compute expected inbreeding coefficients, we use expected and observed heterozygosity, denoted H_e *and* H_o, respectively. Expected heterozygosity corresponds to heterozygosity predicted by Hardy-Weinberg equilibrium. For an infinite population with a given rate of non-random mating α, observed heterozygosity after one generation H_o' is equal to:

$$H_o' = (1 - \alpha)\, H_e + \alpha\, (1 - k) H_o \qquad (3)$$

The various reproductive systems considered here decrease heterozygosity at different rates, *k*, per generation. A proportion *1 - α* of the population reproduces by random mating, producing the expected level of heterozygosity. The remaining fraction, α, does not mate at random, removing a part *k* of previous generation observed heterozygosity H_o. At equilibrium ($H_o' = H_o$), we obtain:

$$H_o = \frac{1-\alpha}{1-(1-k)\alpha} H_e \quad (4)$$

Expected inbreeding coefficients are computed as $F = 1 - H_o/H_e$. Hence:

$$F = \frac{k\,\alpha}{1-(1-k)\alpha} \quad (5)$$

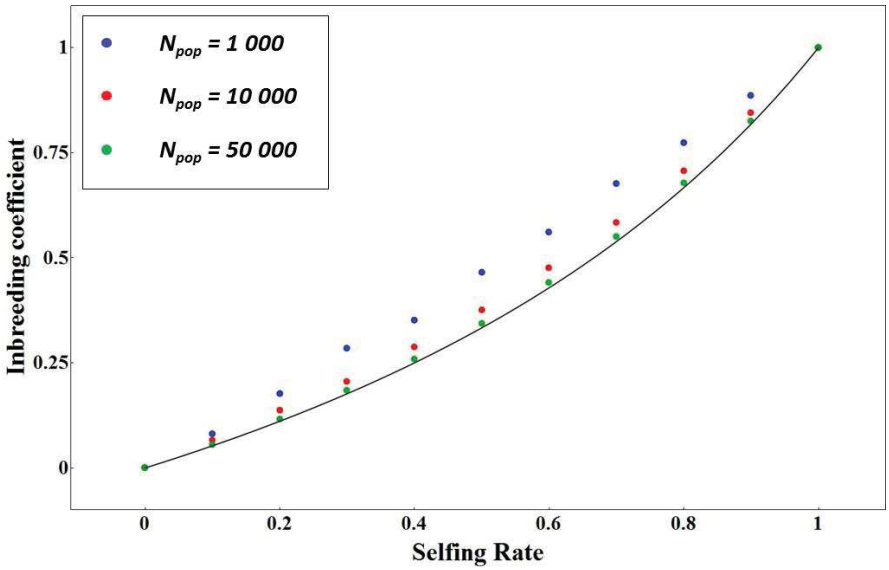


Fig. 2. Expected and observed inbreeding coefficients (y_axis) under neutrality for populations with different rates of self-fertilization (x-axis). Expected inbreeding as computed from Eq. 5, and neglecting inbreeding caused by finite population sizes, is represented as a, black line. Observed inbreeding coefficients are computed from heterozygosity measured in simulated populations. Three population sizes are illustrated : $N_{pop} = 100$ (blue dots), $N_{pop} = 1\,000$ (red dots) and $N_{pop} = 5\,000$ (green dots)

We illustrate on Fig 2 expected (black line, calculated with Eq. 5) and observed (points) inbreeding coefficients in populations of different sizes. To calculate observed inbreeding coefficients, we measure observed heterozygosity and use observed heterozygosity at $\alpha = 0$ as expected heterozygosity. The differences between expected and observed inbreeding coefficients become smaller for bigger populations, as expected from the occurrence of residual inbreeding caused by finite population size.

At the start of simulations, all individuals are homozygotes for both loci. For 1000 generations, mutations on the gene locus occur as explained above. After these burn-in generations, the gene locus is close to selection-mutation equilibrium, and mutations on the enhancer locus begin. In simulations, we measure the frequency of heterozygotes in the population and the rate of enhancer strength escalation. The latter is computed as the slope of mean enhancer log-strength through time (which increases linearly, see details in FYON *et al.* 2015). When investigating the question of enhancer and gene divergence in cases of full Apomixis, full Central Fusion and full Pre-Meiotic Doubling, we follow the mean enhancer log-strength of the stronger enhancer-bearing chromosome (E^+), mean enhancer log-strength of the weaker enhancer-bearing chromosome (E^-), and mean fitness effect of gene alleles on corresponding chromosomes ($\bar{s}^+$ and $\bar{s}^-$). To provide a measurement of allele-specific expression, we calculate and display on Fig. 3.b the log-ratio of homolog enhancers' strengths $Log(E^+/E^-)$. By definition, this log-ratio is always positive.

Results

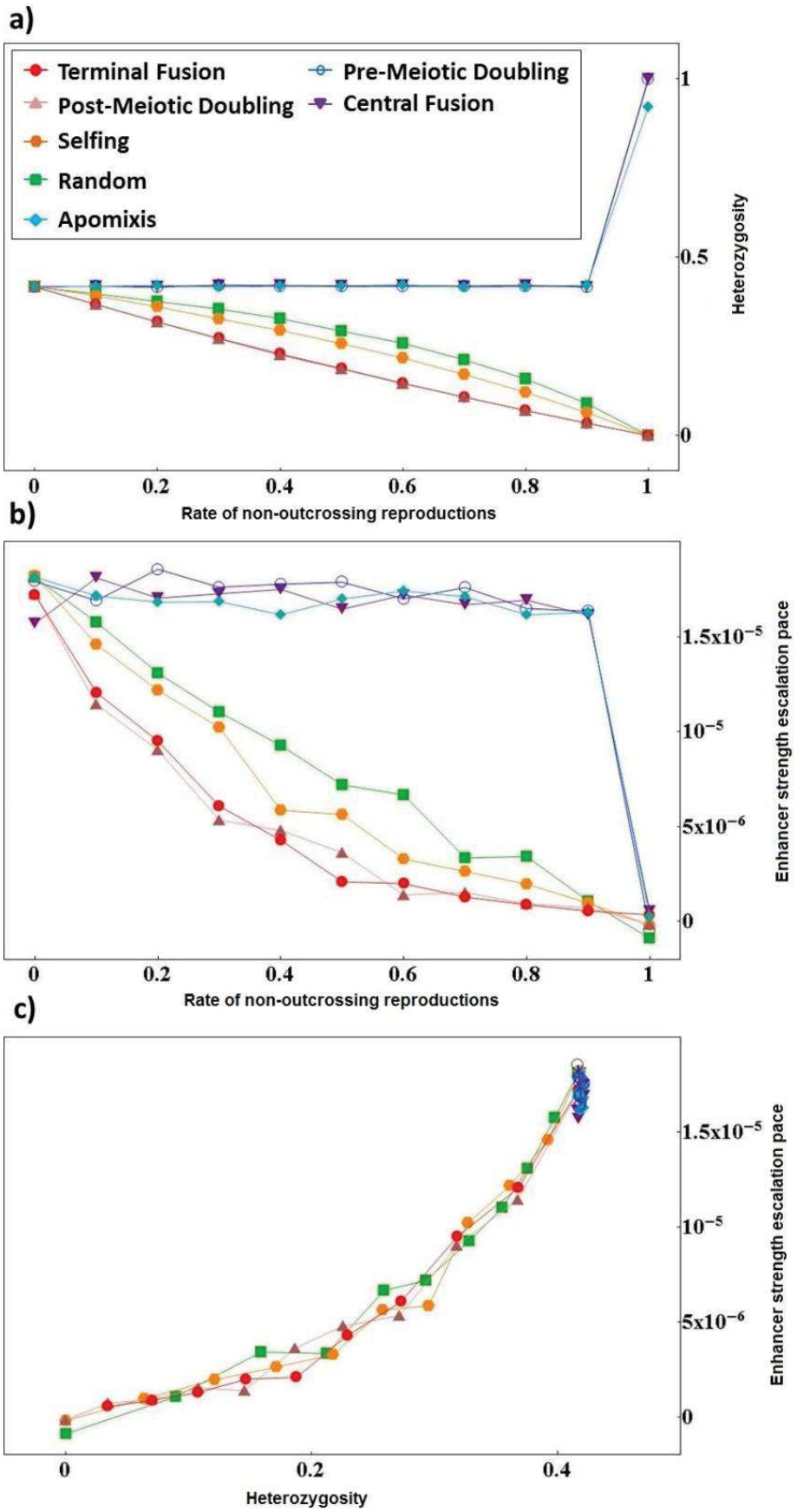
Effect of heterozygosity on the ER process

We first checked the effects of the different reproductive systems on heterozygosity levels. Results are shown on Fig 3.a, plotted against the rate of asexual reproduction per generation (a rate of *0.2* apomixis, for example, indicates that, each generation, 20% of individuals reproduce without meiosis, while 80% of them outcross). We first ignore the recombination between the gene-enhancer pairs and the centromeres, which has no consequence except for terminal and central fusion automixis. For the latter, we start considering that this rate is zero, but this assumption will be relaxed later. First we see, as expected, that apomixis, pre-meiotic doubling and central fusion automixis do not alter heterozygosity compared to fully outcrossing populations (except at low levels of outcrossing, when chromosomes diverge, see below). On the contrary, all other reproductive systems tend to decrease heterozygosity below the Hardy-Weinberg expectation (positive inbreeding coefficient). Heterozygosity is equally lowest, as expected, between reproductive systems entirely eliminating heterozygosity (post meiotic doubling, or terminal fusion automixis with no recombination with centromere). Equilibrium heterozygosity is then gradually higher with reproductive systems reducing offspring heterozygosity by one half (self-fertilization) or one third (random fusion, which also corresponds to terminal and central fusion occurring with probability 1/3 and 2/3, respectively).

We plot on Fig 3.b enhancer strength escalation against the rate of asexual reproduction. Apomixis, pre-meiotic doubling and central fusion give similar results. They exhibit constant escalation rates that correspond to the rate under random mating. The only difference with random mating occurs at very low level of outcrossing. In that case, escalation rates drop precipitously to zero when the outcrossing rate tends to zero. With other mating systems, the rate of escalation gradually reduces with reduced levels of outcrossing, and also tends to zero when outcrossing rate tends to zero. For a given level of outcrossing, escalation rates are higher for random fusion, intermediate for self-fertilization, and lower for post-meiotic doubling / central fusion. Overall the level of heterozygosity maintained in the population by the different mating systems fully predicts escalation rates (except for cases of full apomixis / pre-meiotic doubling / central fusion). Fig 3.c illustrates this relationship by plotting the

escalation rates against heterozygosity in the populations (full apomixis / pre-meiotic doubling / central fusion points are not shown, as they exhibit a specific behavior, which is examined below). This scaling shows clearly that there is no other effect of mating systems once heterozygosity level has been taken into account. Unless outcrossing rate tends to zero, this analysis shows that reproductive systems impact the ER process only by changing heterozygosity levels in the population. Lower heterozygosity levels reduce the pace of enhancer strength escalation.

Fig. 3. Measures of mean heterozygosity and enhancer strength escalation rate in simulated populations depending on various reproductive systems. Simulated populations reproduce partly through outcrossing and partly through other reproductive systems. **a -** Mean heterozygosity in function of the rate of non-outcrossing. **b –** Enhancer strength escalation rate as a function of the rate of non-outcrossing. **c –** Enhancer strength escalation rate against mean heterozygosity. Enhancer strength escalation rate is computed as the slope of the linear regression fitting the mean increase of enhancer log-strength through time. Individuals that do not reproduce through outcrossing reproduce through Terminal Fusion (red circles), Post-Meiotic Doubling (pink upright triangles), Selfing (orange hexagons), Random fusion (green squares), Apomixis (light blue diamonds), Pre-Meiotic Doubling (blue rings) or Central Fusion (purple downright triangles). Enhancer strength escalation rate increases quadratically with mean heterozygosity, except under full clonality (Pre-Meiotic Doubling or Apomixis or Central Fusion with no recombination).


a)
Terminal Fusion
Pre-Meiotic Doubling
Post-Meiotic Doubling
Central Fusion
Selfing
Random
Apomixis
Heterozygosity
0
0.5
1
0
0.2
0.4
0.6
0.8
1
Rate of non-outcrossing reproductions
b)
Enhancer strength escalation pace
1.5x10⁻⁵
10⁻⁵
5x10⁻⁶
0
0
0.2
0.4
0.6
0.8
1
Rate of non-outcrossing reproductions
c)
Enhancer strength escalation pace
1.5x10⁻⁵
10⁻⁵
5x10⁻⁶
0
0
0.2
0.4
Heterozygosity

A new outcome: 'Enhancer Divergence' process

In cases of full apomixis, full pre-meiotic doubling and full central fusion, ER process does not occur. Indeed, we observe maximal levels of heterozygosity, and yet enhancer strength does not escalate. To understand what happens in these cases, we tracked mean enhancer strength on each homologous chromosome separately in every individual. To do so, we define a stronger enhancer-bearing chromosome (*E+* chromosome) and a weaker enhancer bearing chromosome (*E-* chromosome) in every individual, and calculate the population mean enhancer strength for each of these chromosomes. Results for apomixis are plotted on Fig 4, with a case of full apomixis compared to a case of partial apomixis. Fig 4.a shows that these two cases lead to very different results. In partial apomixis, the ER process occurs as usual: all chromosomes increase in strength. In full apomixis however, there is a divergence between chromosomes in every clonal lineage: one chromosome accumulates enhancer strength-increasing mutations, while the other accumulates enhancer strength-decreasing mutations. Consequences on the gene locus can be readily seen in Fig 4.b. We compute $\bar{s}^+$ and $\bar{s}^-$, the average fitness effects of alleles on the gene for E^+ and E^- chromosomes across individuals, respectively. In partial apomixis both $\bar{s}^+$ and $\bar{s}^-$ reach an equilibrium. The two values differ nevertheless since stronger enhancers tend to be purged and associated to less deleterious alleles. However, this difference does not increase through time. In contrast, with full apomixis, $\bar{s}^+$ and $\bar{s}^-$ diverge through time. On *E+* chromosome, the gene locus is well purged from deleterious alleles, as it progressively becomes the only expressed chromosome ($\bar{s}^+$ decreases). On *E-* chromosome, deleterious mutations accumulate, as they are hidden due to low expression ($\bar{s}^-$ increases). Cases of full pre-meiotic doubling and full central fusion give the same results. Overall, in cases of full apomixis, full pre-meiotic doubling and full central fusion with no recombination with the centromeres (clonal lineages), selection on enhancers results in a divergence of enhancer and gene copies. We refer to this process as the 'Enhancer Divergence' (ED) process. This process results in a progressive haploidization of expression, and in the degeneration of the unexpressed gene copy.

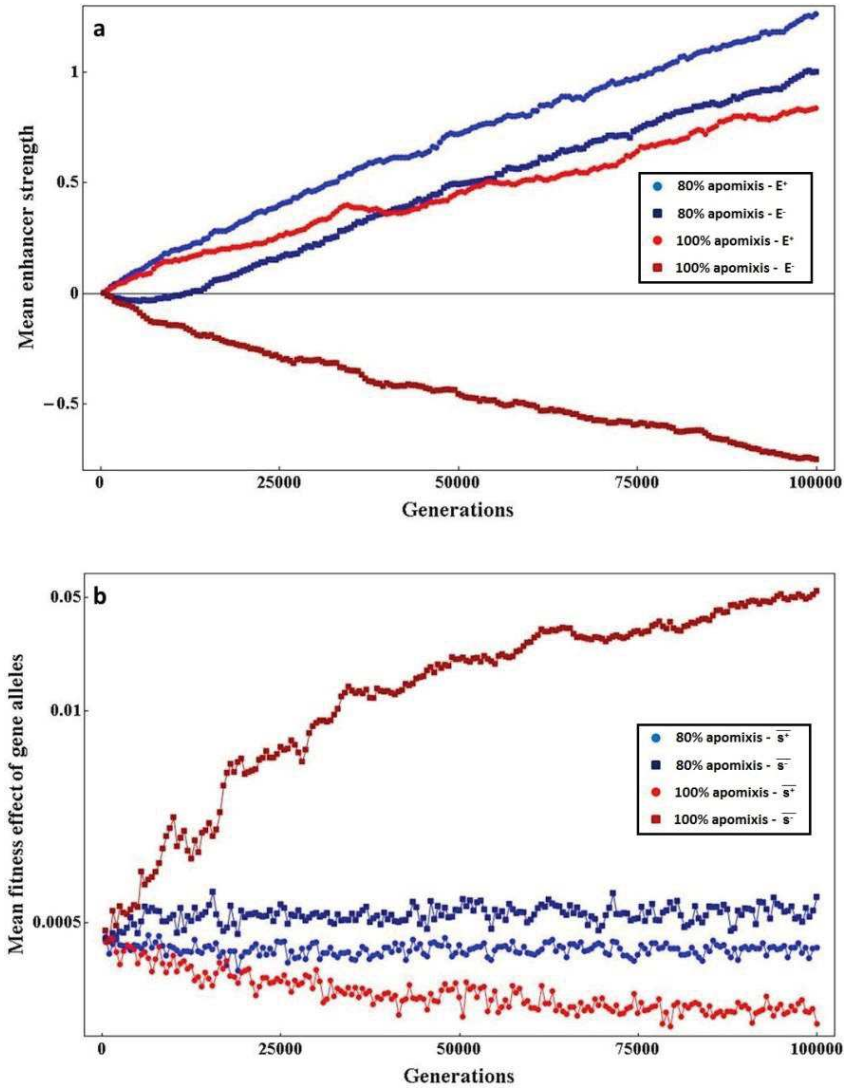


Fig. 4. Evolution through time of mean enhancer strength (**a**) and mean fitness effects of alleles on the gene (**b**) during 100000 generations. Results are shown for chromosomes with stronger enhancers ($\mathbf{E}^+$ and $\bar{s}^+$, circles) and for chromosomes with weaker enhancers ($\mathbf{E}^-$ and $\bar{s}^-$, squares). Results differ depending on whether the population partially reproduces apomictically (80% of reproductions being through apomixis, 20% through outcrossing, in blue) or whether it reproduces entirely through apomixis (100% of reproductions being through apomixis, in red). Under partial apomixis regime, enhancer strengths escalate on both chromosomes. Meanwhile, mean fitness effect of alleles of the gene goes to an equilibrium, with more deleterious alleles on $\mathbf{E}^-$ chromosome since weaker enhancers tend to get preferentially associated with more deleterious alleles. Under total apomixis regime, chromosomes diverge. $\mathbf{E}^+$ chromosomes accumulate stronger enhancer whereas $\mathbf{E}^-$ chromosomes accumulate weaker enhancer. Meanwhile, $\mathbf{E}^+$ chromosomes get purged from deleterious alleles ($\bar{s}^+$ decreases), as they are more expressed than homolog chromosomes. Reversely, $\mathbf{E}^-$ chromosomes accumulate deleterious mutations ($\bar{s}^-$ increases) because they get less and less expressed.

Why do chromosomes diverge in clonal lineages? First it can occur because there is no genetic shuffling. Without recombination, gene conversion or gamete shuffling (through outcrossing or inbreeding), a single mutant enhancer cannot become homozygote in any individual (like for any beneficial mutation in asexuals, see KIRKPATRICK and JENKINS 1989). Enhancer mutations, as well as gene mutations, are restricted to the chromosome on which they arose. In our infinite-allele model, clonal lineages all end up being heterozygous due to enhancer and gene mutations. Hence, selection on the enhancer locus tends to favor individuals expressing less the deleterious alleles that happen to occur on one of the copies of the gene. In other words, genetic association between stronger enhancers and more viable gene allele are selectively favored, as in the ER process, but the result is different. Rather than invading all the population, stronger enhancers concentrate on the chromosome bearing the most favorable gene allele in every individual. The process is self-sustaining: as one gene becomes under-expressed and thus partially hidden from selection, it becomes even more likely to accumulate new deleterious mutations. As one gene accumulates deleterious mutations, enhancer strength-decreasing mutations gets more selected for. This process repeats itself until haploid expression is reached at this particular gene. Note that the process occurs independently for each gene, so that each homolog may maintain expression for different genes. Both homologs are expected to preserve some expressed genes, and none is expected to fully degenerate.

From 'Enhancer Runaway' to 'Enhancer Divergence'

To fully understand the differences between ER and ED processes, we compared enhancer strength escalation between both processes, on a longer timescale than on Fig 4. Fig 5 illustrates this comparison. While ER process ends up in a linear increase of enhancer strength, it is not the case of mean enhancer strength of stronger-enhancer bearing chromosomes in ED process. In cases of clonal lineages, we see that mean enhancer strength of *E+* chromosomes increase logarithmically. After a sufficient amount of time, the marginal increase in enhancer strength becomes negligible. This is because ED process works through the progressive hiding of deleterious alleles on weaker enhancer bearing chromosomes. In our model, extinction of one gene is asymptotic, but the process stops when nearly full extinction (i.e. haploid expression) is achieved.

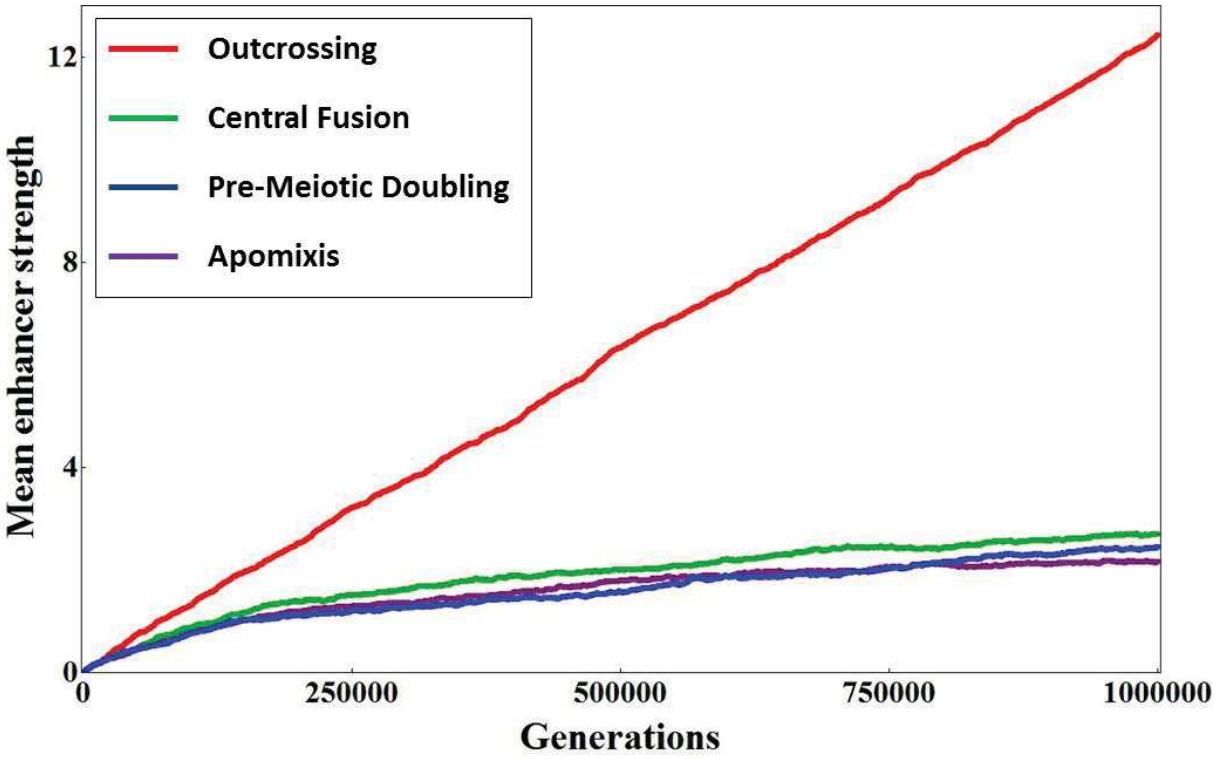
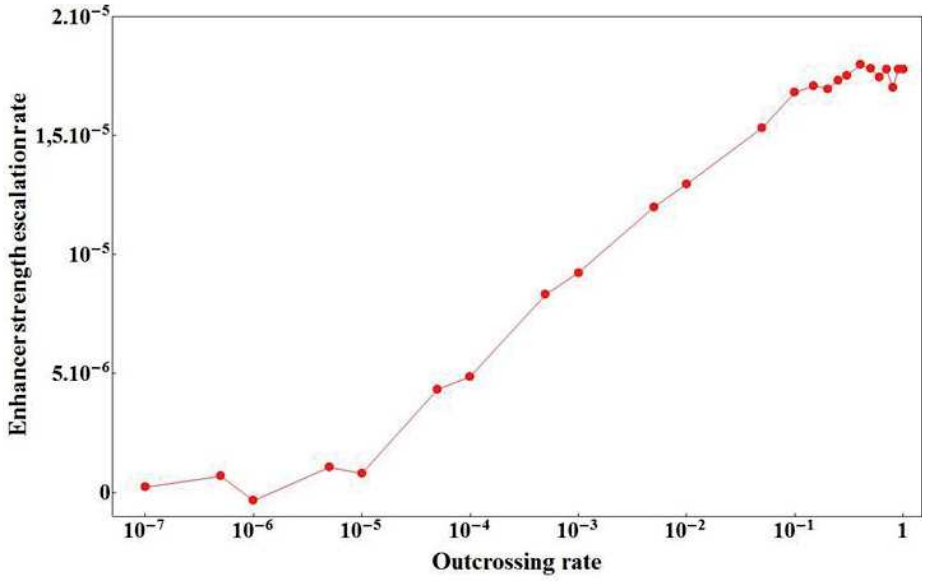
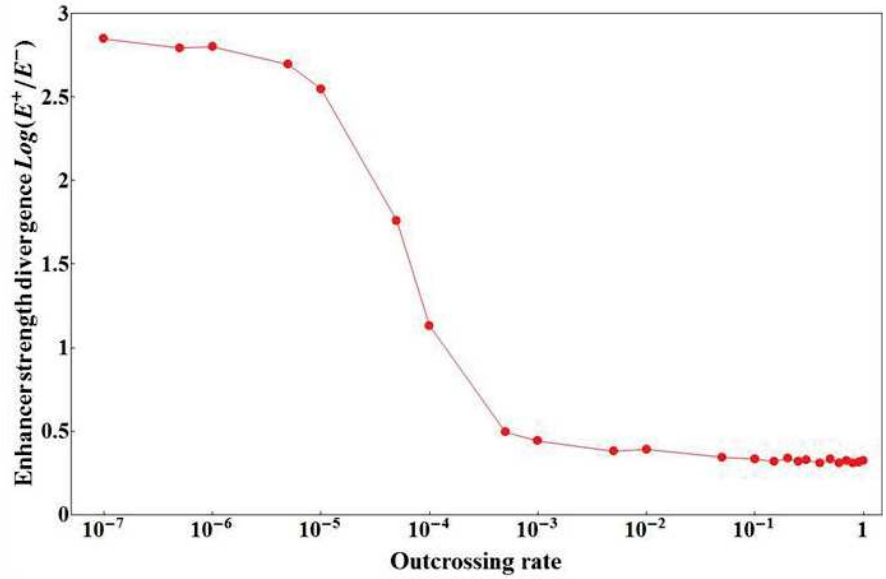


Fig. 5. Comparison of enhancer strength increase with different reproductive system. In Outcrossing populations (red), mean strength of all enhancers increase linearly with time. In clonal lineages (automictic central fusion in green, automictic pre-meiotic doubling in blue, apomixis in purple, with no recombination with the centromeres considered), mean enhancer strength of stronger enhancer bearing chromosomes increase less than linearly.

Fig 3 shows that the transition between the ER and ED processes occurs at low rates of outcrossing. In order to better understand the transition between this two regimes, we plotted on Fig 6 the escalation rate, enhancer strength divergence ratio ($Log(E^{+}/E^{-})$) and heterozygosity of populations with various rates of central fusion automixis without recombination with the centromeres. We see that the transition is progressive, between outcrossing rates of 10^{-5} and 10^{-1}. As outcrossing rate becomes small, escalation rates decrease (the ER process leads to slower escalation at lower levels of heterozygosity, see Fig 3), while heterozygosity and enhancer divergence increase (the ED process becomes stronger as genetic shuffling and recombination decrease). The transition shows that ED process only occurs at very low rates of outcrossing.


Enhancer strength escalation rate
2.10⁻⁵
1,5.10⁻⁵
10⁻⁵
5.10⁻⁶
0
10⁻⁷
10⁻⁶
10⁻⁵
10⁻⁴
10⁻³
10⁻²
10⁻¹
1
Outcrossing rate


Enhancer strength divergence Log(E⁺/E⁻)
3
2.5
2
1.5
1
0.5
0
10⁻⁷
10⁻⁶
10⁻⁵
10⁻⁴
10⁻³
10⁻²
10⁻¹
1
Outcrossing rate

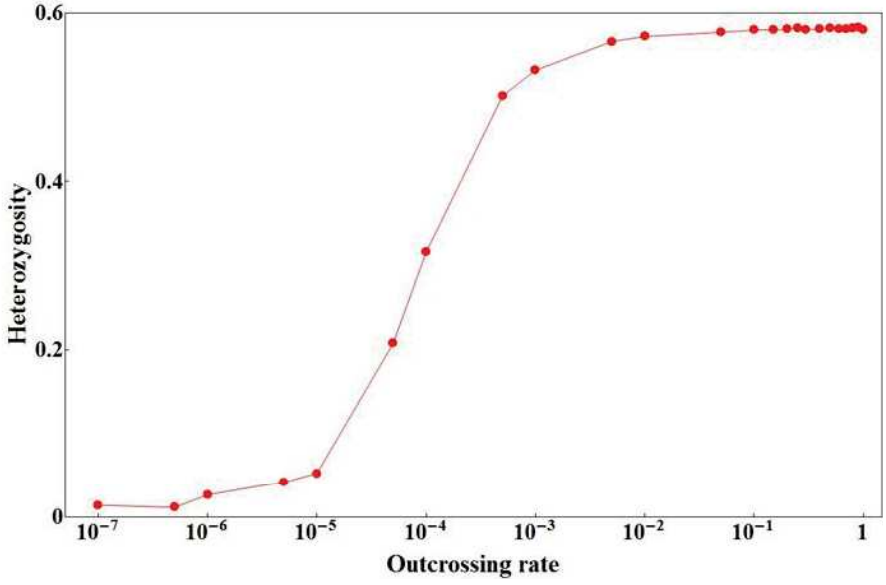

Heterozygosity
0.6
0.4
0.2
0
10⁻⁷
10⁻⁶
10⁻⁵
10⁻⁴
10⁻³
10⁻²
10⁻¹
1
Outcrossing rate

Fig. 6. Transition between *ER process* (high rates of outcrossing) and *ED process* (low rates of outcrossing) in populations reproducing partly through outcrossing, partly through automictic central fusion. Y-axes represent enhancer strength escalation pace (**a**), enhancer strength divergence ratio (**b**) and equilibrium heterozygosity (**c**).

We now consider in more details divergence in the particular case of central fusion automixis. In this reproductive system, the loss of heterozygosity per generation at a given locus depends on the recombination rate with the centromeres. This situation creates a gradient of heterozygosity within each chromosome. While heterozygosity is fully preserved at the centromere (like for apomixis, see results discussed so far), it becomes progressively lost when the recombination rate with the centromere increases. Hence, with full central fusion automixis (no outcrossing at all) we expect enhancer divergence to only occur near centromeres. Fig 7 illustrates this situation (full central fusion with different recombination rates between the enhancer-gene pairs and the centromeres). As expected, chromosome divergence decreases as the recombination with the centromeres increases. Three regimes can be defined on the basis of the results displayed on Fig 7. At low recombination rate, chromosomes diverge and enhancer strength does not escalate as detailed above (ED and no ER). At high recombination rate, no divergence and no strength escalation occur (no ED and no ER), as heterozygosity is not maintained. At intermediate recombination rate, there is some increase of mean enhancer log-strength, and a decrease of chromosome divergence compared to the low recombination regime (little ED and ER).

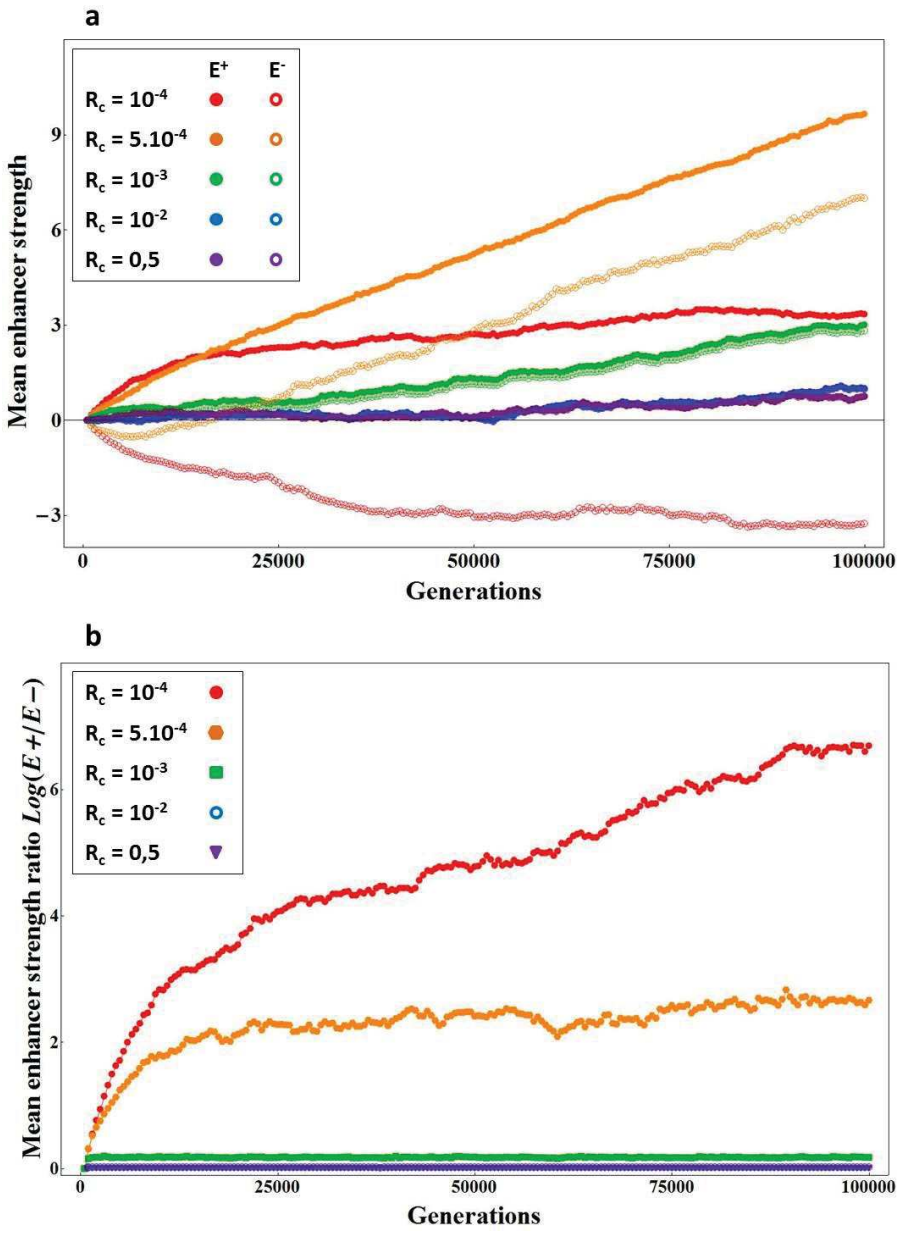


Fig 7. Enhancer runaway and divergence in totally automictic populations reproducing through Central Fusion. **a –** Mean enhancer strength of stronger enhancer bearing (**E**$^{+}$, circles) and weaker enhancer bearing (**E**$^{-}$, rings) chromosomes. Results are given for various recombination rate between enhancer-gene loci and centromeres (R_c): $R_c = 10^{-4}$ (red), $R_c = 5.10^{-4}$ (orange), $R_c = 10^{-3}$ (green), $R_c = 10^{-2}$ (blue), and $R_c = 0,5$ (purple). Results heavily depend on R_c values, with three regimes: divergence without escalation ($R_c = 10^{-4}$), small divergence and escalation ($R_c = 5.10^{-4}$), no divergence and no escalation ($R_c = 0,5$). **b –** Mean enhancer strength divergence between homologous chromosomes. Divergence is calculated as the decimal logarithm of **E**$^{+}$ mean enhancer strength over **E**$^{-}$ mean enhancer strength. Same values of R_c are used: $R_c = 10^{-4}$ (red circles), $R_c = 5.10^{-4}$ (orange hexagons), $R_c = 10^{-3}$ (green squares), $R_c = 10^{-2}$ (blue rings), and $R_c = 0,5$ (purple downright triangles). Divergence between homolog chromosomes plumets as R_c increases.

Discussion

The two possible outcomes of indirect selection on enhancers

The enhancer runaway process is very general and should occur on every gene in every diploid species (FYON *et al.* 2015). It should not occur however for enhancers located too far from their gene or for populations displaying low heterozygosity. As we demonstrate here, the amount of heterozygosity is a crucial determinant of the rate of the ER process. ER rate increases with heterozygosity level. This increase is also faster than linear, as expected from the fact that both the enhancer and gene loci need to be heterozygous for the process to operate. We also uncover that another process can take place when the mating system is very close to (or equivalent to) full clonality. In this case, homolog genes and enhancers diverge. This process that we refer to as the Enhancer Divergence (ED) process, leads to haploidization of expression. One purged gene copy is well expressed while the other accumulates deleterious mutations and stops being expressed. This process does not necessarily lead to the degeneration of one chromosome as the two homologs each preserve a distinct subset of active genes in diploid clonal lineages.

Why does the same indirect selection pressure lead to two distinct evolutionary processes? This is because, for the ER process to occur, a stronger mutant enhancer needs to be able to invade the population. In clonal lineages, when descendants are identical to their parents (except for mutations), this invasion is impossible. Without recombination or gene conversion, regulatory mutants stay confined to the chromosome lineage on which they appear. The favorable association between stronger enhancers and genes with fewer deleterious mutations cannot lead to the invasion of the whole population by the stronger enhancers. It rather leads, for each gene concerned, to the accumulation of stronger enhancers and non-deleterious gene allele on one chromosome, and weaker enhancers and deleterious gene alleles on the other homolog chromosome. Which homologous copy of the gene becomes silenced entirely depends on the initial stochasticity of occurrence of mutations on the gene and its regulatory region.

Divergence patterns in some modes of automixis, such as central fusion or genetically equivalent systems (e.g. suppression of meiosis I), depend on the recombination rate between the enhancer-gene pairs and the centromere. Three regimes can be identified. At

low recombination, ED, but not ER, process occurs. At intermediate recombination, both ER and ED processes occur. At high recombination, ER and ED do not significantly occur. The reason for this pattern is that (i) ED process becomes negligible as recombination creates homozygotes and reduces divergence among regulatory regions, (ii) ER process requires some recombination to occur (for the stronger enhancer to invade all chromosomes), but not too much (ER process only works when heterozygotes are present). Note that considering a multilocus model may quantitatively alter the recombination rate at which those transitions occur. Indeed, when there is a rare recombination event leading to a loss-of-heterozygosity, all loci that are distal to the crossing over position (i.e between this position and the telomere) become homozygotes. Some of these loci will become homozygote for the beneficial association (stronger enhancer and purged gene copy), but some will become homozygote for the other association between weaker enhancers and deleterious gene copies. Hence, the closer the crossing over position is to the centromere, the less likely it is that it will lead to viable offspring. In other words, lineages that survive are those that experienced few recombination events, and recombination events that are located distally from the diverging enhancer-gene pairs. This low effective recombination rate is then likely to promote divergence at larger genetic distance (from the centromere) than in the single locus model. We may even expect divergence to progressively spread from the centromere through time. Independently, recombination suppression may also evolve simply to prevent loss-of-heterozygosity and the exposition of deleterious mutations in homozygous state.

This dynamic process of Enhancer divergence, like all models of allele divergence in asexuals, relies on the fact that homologous chromosomes remain 'isolated' from each other. This isolation can be impacted by several processes that we did not include in our model (as for Meselson effect, see discussion in e.g. Butlin 2002). Recombination with the centromeres in central fusion breaks this genetic isolation, and that is why it prevents ED process. Gene conversion, which is not taken into account in our model, is similarly able to break this isolation. Clonal species harboring high rates of gene conversion, might escape ED process. Also, mitotic recombination may alter chromosome isolation in apomictic species. A better assessment of actual chromosome isolation in asexual species is necessary to better envision the importance of ED process.

Predictions

The different outcomes of ED and ER processes lead to different predictions that could be empirically investigated. Concerning the ER process, we predict that enhancers will be generally stronger in populations with higher heterozygosity levels. This may lead to a variety of tests. For example, this could be investigated by measuring allele-specific expression in F1 hybrids between species differing by their reproductive systems. In such test, as all *trans*-acting regulators from each parents are shared in the hybrid, any difference in transcript abundance between genes coming from each parents are due to differences in *cis*-acting regulators. Because the rate of ER process increases with heterozygosity levels, we expect genes inherited from the outcrossing species to be on average more expressed in the hybrid. Expression patterns in this kind of hybrid have been investigated by HE and colleagues (2012). They crossed outcrossing plants *Arabidopsis lyrata* (as male) with selfing *Arabidopsis thaliana* (as female), and found that, in 90% of the genes displaying allele-specific expression, the *lyrata* genome was preferentially expressed. However, this may be also caused by a sex-of-origin (rather than a species-of-origin) effect as the study lacks reciprocal crosses due to technical difficulties. A similar study was performed by STEIGE and colleagues (2015) using outcrossing *Capsella grandiflora* (as female) and selfing *Capsella rubella* (as male). In this case, results were in the opposite direction (*C. rubella* genes tended to be preferentially expressed in the hybrid). But again, sex-of-origin effects cannot be ruled out. Finding a model system where the two different directions of the cross could be performed would be particularly insightful.

More generally, this prediction could be tested comparing strength of enhancers between populations with different heterozygosity levels. Reproductive systems are not the only factor influencing heterozygosity. Other factors include: mutation, migration, non-random mating, selection, recombination, gene conversion and population size. Such factors act in concert. Rather than disentangling the effects of each factor on heterozygosity and then on enhancer strength, one may simply assess heterozygosity and enhancer strengths in different populations and see if they correlate positively. The detection of such a correlation would be an important clue towards confirmation of the occurrence of ER process and its impact on expression regulation evolution. To circumvent the problem of obtaining hybrids,

it may be possible to use models calibrating enhancer strength from interactions with different *trans*-acting regulators as with *STARR*-seq methods (MUERDTER *et al.* 2015).

In some asexual species displaying very high levels of heterozygosity (apomixis, pre-meiotic doubling, central fusion without recombination with the centromeres), haploidization of expression should evolve. This prediction may be modulated depending on the degree of genetic isolation between homolog chromosomes. With central fusion automixis, we predict that regions close to the centromeres should exhibit haploid or nearly haploid expression. On the contrary, away from centromeres, such divergence should not occur and the rate of ER process should also be small, owing to the low level of heterozygosity. Such genome would therefore offer a very strong and simultaneous test of both ER and ED processes if hybrids can be somehow obtained between a parthenogenetic and a related sexual species. More generally, in parthenogenetic lineages, haploidization should evolve whenever chromosomes become isolated enough due to the absence of sex, recombination and gene conversion. Haploid expression in diploid species can be revealed using allele-specific expression assessing techniques, like RNA-seq or micro / oligo arrays. ED process should create a signal of large and widespread allele-specific expression. In central fusion species, allele-specific expression should decrease with increased genetic distance from the centromeres.

Overall, we showed that the evolution of regulatory regions strongly depends on reproductive systems, beyond the stochastic effects (mutation, drift, selective interference) considered in classical population genetics models. These selective effects result from competition for expression of cis-regulatory regions. They lead to a blend of runaway and divergence depending on the rate of sex. These selective effects are also entirely unrelated to optimization of expression levels due to organismal-level selection.

References

ABZHANOV A., PROTAS M., GRANT B. R., GRANT P. R., TABIN C. J., 2004 Bmp4 and morphological variation of beaks in Darwin's finches. Science **305**: 1462–5.

ASHER J. H., 1970 Parthenogenesis and Genetic Variability. II. One-Locus Models for Various Diploid Populations. Genetics **66**: 369–391.

BUTLIN R., 2002 The costs and benefits of sex: new insights from old asexual lineages. Nat Rev Genet **3**: 311–317.

CARROLL S. B., 2005 Evolution at Two Levels: On Genes and Form. PLoS Biol. **3**: e245.

CARROLL S. B., 2008 Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. Cell **134**: 25–36.

COOLON J. D., MCMANUS C. J., STEVENSON K. R., GRAVELEY B. R., WITTKOPP P. J., 2014 Tempo and mode of regulatory evolution in Drosophila. Genome Res. **24**: 797–808.

COOPER T. F., 2003 Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. **100**: 1072–1077.

FEREA T. L., BOTSTEIN D., BROWN P. O., ROSENZWEIG R. F., 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc. Natl. Acad. Sci. U. S. A. **96**: 9721–9726.

FORCE A., CRESKO W. A., PICKETT F. B., PROULX S. R., AMEMIYA C., LYNCH M., 2005 The Origin of Subfunctions and Modular Gene Regulation. Genetics **170**: 433–446.

FYON F., CAILLEAU A., LENORMAND T., 2015 Enhancer Runaway and the Evolution of Diploid Gene Expression. PLoS Genet **11**: e1005665.

HE F., ZHANG X., HU J., TURCK F., DONG X., GOEBEL U., BOREVITZ J., MEAUX J. DE, 2012 Genome-wide Analysis of Cis-regulatory Divergence between Species in the Arabidopsis Genus. Mol. Biol. Evol. **29**: 3385–3395.

HOEKSTRA H. E., COYNE J. A., 2007 The locus of evolution: evo devo and the genetics of adaptation. Evolution (N. Y). **61**: 995–1016.

KHAITOVICH P., WEISS G., LACHMANN M., HELLMANN I., ENARD W., MUETZEL B., WIRKNER U., ANSORGE W., PÄÄBO S., 2004 A Neutral Model of Transcriptome Evolution. PLoS Biol. **2**: e132.

KIRKPATRICK M., JENKINS C. D., 1989 Genetic segregation and the maintenance of sexual reproduction. Nature **339**: 300–301.

KUO D., LICON K., BANDYOPADHYAY S., CHUANG R., LUO C., CATALANA J., RAVASI T., TAN K., IDEKER T., 2010 Coevolution within a transcriptional network by compensatory trans and cis mutations. Genome Res.: 1–7.

LENORMAND T., ROZE D., ROUSSET F., 2015 Stochasticity in evolution. Trends Ecol. Evol. **24**: 157–165.

LYNCH M., 2007 The evolution of genetic networks by non-adaptive processes. Nat Rev Genet **8**: 803–813.

LYNCH V. J., WAGNER G. P., 2008 Resurrecting the Role of Transcription Factor Change in Developmental Evolution. Evolution (N. Y). **62**: 2131–2154.

MANNA F., GALLET R., MARTIN G., LENORMAND T., 2012 The high-throughput yeast deletion fitness data and the theories of dominance. J. Evol. Biol. **25**: 892–903.

MUERDTER F., BORYŃ Ł. M., ARNOLD C. D., 2015 STARR-seq — Principles and applications. Genomics **106**: 145–150.

NOUGUÉ O., RODE N. O., JABBOUR-ZAHAB R., SÉGARD A., CHEVIN L. -M., HAAG C. R., LENORMAND T., 2015 Automixis in Artemia: solving a century-old controversy. J. Evol. Biol. **28**: 2337–2348.

OLEKSIAK M. F., CHURCHILL G. A., CRAWFORD D. L., 2002 Variation in gene expression within and among natural populations. Nat. Genet. **32**: 261–6.

RAYMOND M., CHEVILLON C., GUILLEMAUD T., LENORMAND T., PASTEUR N., 1998 An overview of the evolution of overproduced esterases in the mosquito Culex pipiens. Philos. Trans. R. Soc. B Biol. Sci. **353**: 1707–1711.

SCHÖN I., MARTENS K., DIJK P. VAN, 2009 *Lost Sex The Evolutionary Biology of Parthenogenesis*. Springer, Dordrecht.

SHAPIRO M. D., MARKS M. E., PEICHEL C. L., BLACKMAN B. K., NERENG K. S., JÓNSSON B., SCHLUTER D., KINGSLEY D. M., 2004 Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature **428**: 717–23.

STEIGE K. A., REIMEGÅRD J., KOENIG D., SCOFIELD D. G., SLOTTE T., 2015 Cis-Regulatory Changes

Associated with a Recent Mating System Shift and Floral Adaptation in Capsella. Mol. Biol. Evol. **32**: 2501–2514.

SUOMOLAINEN E., LOKI J., 1987 *Cytology and Evolution in Parthenogenesis*. CRC Press, Boca Raton, FL.

TAUTZ D., 2000 Evolution of transcriptional regulation. Curr. Opin. Genet. Dev. **10**: 575–579.

WEIRAUCH M. T., HUGHES T. R., 2010 Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. **26**: 66–74.

WITTKOPP P. J., HAERUM B. K., CLARK A. G., 2004 Evolutionary changes in cis and trans gene regulation. Nature **430**: 85–8.

WITTKOPP P. J., KALAY G., 2012 Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. **13**: 59–69.

WRAY G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. **8**: 206–216.

WRAY G. A., HAHN M. W., ABOUHEIF E., BALHOFF J. P., PIZER M., ROCKMAN M. V, ROMANO L. A., 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–419.

YANAI I., GRAUR D., OPHIR R., 2004 Incongruent Expression Profiles between Human and Mouse Orthologous Genes Suggest Widespread Neutral Evolution of Transcription Control. Omi. A J. Integr. Biol. **8**: 15–24.

La sélection indirecte sur les séquences régulatrices peut entraîner deux types de processus. Tout d'abord, si les chromosomes homologues ne sont pas isolés génétiquement, il y a l'*ER process* : les promoteurs forts envahissent parce qu'ils sont associés à des allèles viables du gène. Ce processus a lieu primitivement chez les hétérozygotes, et donc dépend fortement, entre autres, de l'hétérozygotie de la population. Le second processus a lieu lorsque les chromosomes homologues sont isolés génétiquement, comme dans le cas de lignées clonales peu ou pas recombinantes. Dans ce cas, les promoteurs forts ne peuvent pas envahir l'ensemble de la population, et les individus demeurent hétérozygotes. Les associations génétiques entre promoteurs forts et allèles viables se concentrent alors sur un chromosome, tandis que les associations génétiques entre promoteurs faibles et allèles délétères se concentrent sur le chromosome homologue. La divergence des promoteurs des chromosomes homologues (*ED process*) est à l'origine d'une haploïdisation progressive de l'expression des gènes concernés.

Les cas étudiés ne sont pas les seuls exemples de chromosomes isolés génétiquement. Un autre exemple classique est le cas des chromosomes sexuels qui ne recombinent plus. Nous avons donc étudié la conséquence de la sélection indirecte sur les séquences régulatrices dans un cas où les gènes concernés se trouvent sur des chromosomes sexuels. Dans ce cas, le processus de divergence des promoteurs a également lieu entre le X et le Y (ou le W et le Z). De façon intéressante, c'est toujours le chromosome Y qui accumule des promoteurs faibles et des mutations délétères, alors que c'est toujours le chromosome X qui accumule des promoteurs forts associés à des allèles viables. Il semble ainsi que le processus de divergence des promoteurs soit en mesure d'expliquer en partie l'évolution des jeunes chromosomes sexuels, après qu'ils aient cessé de recombiner : accumulation d'allèles délétères et extinction de l'expression sur le chromosome Y, augmentation de l'expression du chromosome X. Des compensations de dosage chez la femelle peuvent avoir lieu si l'augmentation de l'expression des chromosomes X n'est pas limitée aux mâles.

Bien que la sélection indirecte à l'origine du processus soit faible, le processus est très général et pourrait contribuer à expliquer la dégénérescence des chromosomes Y et W. Il est surtout efficace en grandes populations, là où les théories actuelles, basées sur des effets d'interférence sélective, ont du mal à expliquer l'évolution des chromosomes sexuels. De façon intéressante, notre processus associe simultanément extinction et dégénérescence du

chromosome Y, alors que les théories actuelles ont plutôt tendance à présenter l'extinction comme un moyen de compenser la dégénérescence, ou au contraire la dégénérescence comme résultant de l'extinction. La surexpression du chromosome X est également concomitante dans notre processus, alors qu'elle est souvent vu comme une compensation de dosage postérieure à l'extinction du chromosome Y.

Paper in preparation

Chapitre IV

Enhancer Divergence in Sex Chromosomes: a new theory for Y chromosome erosion

Authors : Frédéric Fyon, Thomas Lenormand

UMR 5175 CEFE, CNRS - Université Montpellier - Université P. Valéry - EPHE, 1919 route de Mende 34293 Montpellier Cedex 5, France

Abstract

It is usually thought that Y (or W) chromosome degeneration is the result of less efficient selection on non-recombining regions. Degeneration of the Y would lead to non-functional copies of genes on this chromosome and an overall decrease in expression of functional proteins in XY (WZ) individuals (only the copy of the gene on the X being functional). There may also be an adaptive decrease of Y expression / increase of X expression in order to hide degenerate Y gene copies. This situation can cause an expression imbalance with autosomes in XX or XY individuals. For dosage-sensitive genes, expression imbalance is deleterious. In these cases, various dosage compensation systems have evolved to balance expression of autosome and sex chromosome genes.

Here, we argue that another process, the *Enhancer Divergence process*, participates to the accumulation of deleterious mutations on Y (W) chromosomes, and is also causing the decrease of Y (W) expression and the simultaneous increase of X (Z) expression. This process of Y chromosome degeneration and associated expression changes is initiated once recombination is suppressed between sex chromosomes. The theory presented here does not involve selective interference as proposed in current theories. This new theory may better account for the evolution of sex chromosomes in big populations. It can work in concert with the accumulation of deleterious mutations on the Y caused by selective interference.

Introduction

Sex chromosomes are a particularly interesting feature of eukaryotic genomes. *XX/XY* and *ZZ/ZW* sex determination systems share important and convergent similarities (ELLEGREN 2011). Both of them involve a chromosome that is always heterozygous (*Y* and *W*), and present in only one gender. All or part of this chromosome has stopped recombining. It is considerably degenerate in many species, and only a fraction of the ancestral Y/W chromosome remains functional. There is a consensus on the early stages of Y chromosome evolution (from now on, we will only refer to X/Y chromosomes, but everything applies to Z/W chromosomes). The first step is the segregation of male-determining genes on the Y chromosome, and female-determining genes on the X chromosome (BULL 1983; CHARLESWORTH 1996; BACHTROG 2006b). Subsequently, recombination stops between sexually antagonistic loci and the sex-determination locus to ensure that sexually antagonistic alleles segregate in their respective sex (CHARLESWORTH and CHARLESWORTH 1980; RICE 1987a; LENORMAND 2003). Recombination may also stop at some locations where different chromosome rearrangements have spread on X and Y chromosomes, in order to prevent meiotic pairing failure due to lack of structural or conformational homology (JABLONKA and LAMB 1990; IRONSIDE 2010). The breakdown of recombination often spreads over the whole chromosomes on successive 'strata' (MING *et al.* 2007).

The next step has been more controversial. Several theories have been proposed to account for the degeneration of the Y chromosome following recombination suppression. If some, as Muller's theory of recessive deleterious mutation accumulation, have been ruled out (CHARLESWORTH 1978), many of them are still discussed in terms of quantitative importance (CHARLESWORTH and CHARLESWORTH 2000; BACHTROG 2006b).

The first hypothesis, 'Muller's Ratchet', was originally developed in the context of the evolution of sex/recombination (MULLER 1964; FELSENSTEIN 1974). It was then suggested as a mechanism leading to Y degeneration following suppression of recombination (CHARLESWORTH 1978). In this process, non-recombining chromosomes accumulate deleterious mutations: the ratchet clicks because the class of chromosomes bearing the fewest number of deleterious alleles is stochastically lost every once in a while. In absence of recombination and reverse mutations, this is irreversible, as there is no process that can restore

chromosomes with fewer deleterious mutations. As a stochastic process, it strongly relies on population size (CHARLESWORTH 1996). Simulations have shown that Muller's Ratchet alone induces exceedingly slow deleterious mutation accumulation in large populations (CHARLESWORTH 1996), except for mildly deleterious mutations (GORDO and CHARLESWORTH 2000).

The second has been introduced by Rice and is referred to as the 'Hitchhiking Effect' (MAYNARD-SMITH and HAIGH 1974). It proposes that, when a beneficial mutation occurs on the Y chromosome, it is unlikely to occur on a deleterious mutation-free chromosome. As there is no recombination, the sweep of the beneficial mutation will bring many mildly deleterious mutations along (RICE 1987b). Consequently, the fixation of the beneficial mutation causes the fixation of a Y chromosome that probably has more deleterious mutations than the class of Y chromosomes that previously had the fewest number of mutations. However, this process is limited to cases where the beneficial mutation arises in genetic background with a limited number of deleterious mutations, such that their total effect does not completely offset the effect of the beneficial mutation (MANNING and THOMPSON 1984). This is a strong limitation, as beneficial mutations tend to be rare and with limited beneficial effects. This process has thus probably little importance on Y degeneration (CHARLESWORTH 1996).

The third theory proposes that "background selection" tends to reduce genetic diversity at linked loci (CHARLESWORTH 1994). Consequently, on non-recombining Y chromosomes, beneficial mutations have a lower probability of fixation than on the X chromosome, while deleterious mutations have a higher probability of fixation (BIRKY and WALSH 1988). Non-recombining regions indeed suffer from the Hill-Robertson effect (HILL and ROBERTSON 1966; FELSENSTEIN 1974): linkage between selected loci reduces the efficacy of selection.

All of these theories actually take a similar view on Y erosion. They use similar models to explain that, due to the absence of recombination on the Y chromosome, beneficial mutations would less probably go to fixation while deleterious mutations would tend to accumulate. The idea is that in non-recombining regions, selection is less efficient, the effective population size drops, such that beneficial mutations are more often lost, and deleterious mutations are more often fixed. The result is that the fitness of the Y chromosome would progressively lag behind the fitness of the X chromosome (CHARLESWORTH

1978, 1996; RICE 1987b). This degeneration, by accumulation of deleterious mutations can lead to non-functional or non-expressed alleles on the Y. It concerns all functional sequences (coding or regulatory). However, the accumulation of deleterious mutations in Y chromosome coding sequences may also trigger the adaptive silencing of expression of (deleterious) alleles on the Y (met by decreasing Y genes' expression and/or by increasing X genes' expression) (ORR and KIM 1998). Hence, the extent to which regulatory regions evolve by the accumulation of beneficial or deleterious mutations is unclear.

Y silencing, in turn, creates a situation of imbalance in relative expression of sex chromosome genes and autosome genes. Dosage compensation is thought to evolve to correct for this imbalance and insure similar expression levels of sex chromosome genes and autosome genes (DISTECHE 2012; ERCAN 2015). However, this situation concerns a minority of dosage-sensitive genes, for which such imbalance is deleterious (PESSIA *et al.* 2012).Timeline goes as: (1) breakdown of recombination on Y chromosome, (2) degeneration of Y-linked genes – Y-fitness lags behind X-fitness, (3) (optionally) adaptive silencing of Y-linked genes and (4) evolution of dosage compensation for dosage-sensitive genes.

A variant of this theory has been more recently proposed: it proposes that the accumulation of deleterious alleles in regulatory sequences leads to the extinction of Y gene expression (BACHTROG 2006a). Y genes being hidden, they would then neutrally accumulate deleterious mutations and degenerate (ZHOU and BACHTROG 2012). This reverses the timeline just mentioned by switching steps (2) and (3). This theory, however, does not explain well why only expression-decreasing mutations would accumulate (expression-increasing mutations also are deleterious) and why only regulatory regions are hit (many deleterious mutations also occur on coding sequences).

In this paper, we propose a new theory to explain sex chromosome evolution. The main difference with previous theories is that it does not involve selective interference. Also, in this theory, Y degeneration and silencing do not follow each other (in whatever order), but occur at the same time. Contrarily to most of sex chromosome evolution models, we do not assume multiple gene loci, experiencing recurrent mutations with no epistasis. We focus on a simple model with one gene and one *cis*-regulator. We are interested in *cis*-regulators because they impact the expression of the gene copy located on the same chromosome only

(WRAY *et al.* 2003). In previous papers, we showed that such a simple system may undergo, if the recombination rate between the gene and the enhancer is sufficiently low, an Enhancer Runaway (ER) process (FYON *et al.* 2015). During this process, enhancers compete for expression, which leads to a progressive increase in mean enhancer strength (ability of enhancers to promote transcription). This is due to the fact that a stronger enhancer tends to express more its associated gene copy than the homologous enhancer. This gene copy is thus more exposed to selection. Consequently, stronger enhancers tend to be associated to gene copies better purged from deleterious mutations. Stronger enhancers get associated with better genetic background, and invade in a hitchhiking process.

Competition for expression between homologous enhancers does not usually lead to chromosome divergence: enhancers on both chromosomes tend to increase in strength. Yet, in Chapter III, we saw that enhancer divergence may occur in clonal lineages. Simulations showed that in each clonal lineage, one chromosome evolves ever-stronger enhancers while the other chromosome evolves ever-weaker enhancers. This is due to genetic isolation between homologous chromosomes: mutations stay confined to the genetic context where the mutation appeared. A stronger enhancer cannot invade the whole population, as it cannot "jump" to the homologous chromosome. When homologous chromosomes are isolated, all clonal lineages become heterozygous. Fitness optimization then consists in associating deleterious mutations with weaker enhancers on one chromosome, and viable gene copy and stronger enhancers on the other chromosome. We referred to this process as Enhancer Divergence (ED). It is important to note that, for each gene, the chromosome that keeps or loses expression varies, depending on which chromosome is initially hit by mutations in the gene and its regulatory regions. Overall, both chromosomes are eventually still functional, but every gene ends up having haploid expression.

Y chromosome transmission shares many similarities with clonal transmission. We thus investigated the impact of the competition for expression between *cis*-regulators in the case of sex chromosomes. We show that the ED process causes a fast degeneration of genes on the Y chromosome. It also causes an increase in X chromosome gene expression, and a decrease in Y chromosome gene expression. Two cases may arise. In the first case, X chromosome upregulation is limited to males, and proportional to Y chromosome downregulation. Expression level dosage with autosomes is maintained in males and

females. In the second case, X upregulation concerns male and female X chromosomes. Dosage with autosomes is maintained in males, but biased upwards in females. If the gene considered is dosage-sensitive, dosage compensation evolves to restore expression level dosage in females. Overall, we propose that an alternative timeline for sex chromosome evolution could very possibly be: (1) breakdown of recombination on the Y; (2) building of genetic associations: stronger enhancers and viable gene alleles on X chromosomes, weaker enhancers and deleterious gene alleles on Y chromosomes; (3) X chromosomes get upregulated, and Y chromosomes get downregulated and degenerate; (4) dosage compensation in females for dosage-sensitive genes if X upregulation is not restricted to males. Contrarily to current theories, this scenario is particularly efficient, especially in large populations, as it does not rely on selective interference.

Model

To gain insights into the evolution of sex chromosome enhancer strength, we used stochastic simulations of a multi-locus individual-based model. In this model, we consider N_{pop} diploid individuals, half males (*XY*) and half females (*XX*). We suppose that X and Y chromosomes do not recombine. We consider a gene (locus **A**) and its enhancer (locus **E**) located on sex-chromosomes. Enhancers and genes recombine at a rate R_{EA} (in females only). We also consider a modifier of expression **M** (for potential dosage compensation) that recombines freely with the two other loci (the order of loci is **M-A-E**). The life cycle is diploid selection, meiosis with recombination, mutation and syngamy.

Fitness of individuals depends on the gene and enhancer alleles. We note, for an individual *i*, the fitness effects of alleles at locus **A** (the gene) $s_{1,i}$ and $s_{2,i}$ ($s_{1,i} > s_{2,i}$). The fitness of this individual (W_i^A) caused by the presence of deleterious mutations on the gene is calculated as:

$$W_i^A = 1 - s_{2,i} - h_i\left(s_{1,i} - s_{2,i}\right), \quad (1)$$

where h_i is the dominance coefficient of the most deleterious allele among the two alleles carried by individual *i*. This dominance coefficient depends on the relative expression of the two alleles: the less an allele is expressed, the less it impacts individual's fitness, and thus the lower is its dominance coefficient. If we note $e_{1,i}$ and $e_{2,i}$ the strengths of enhancer alleles associated with gene alleles of fitness effects $s_{1,i}$ and $s_{2,i}$ respectively, then the dominance coefficient of the most deleterious allele in individual *i* (h_i) is calculated as:

$$h_i = \left(\frac{e_{1,i}}{e_{1,i} + e_{2,i}}\right)^{-\frac{Log(h)}{Log(2)}} \quad (2)$$

In Eq. 2, *h* is a default dominance coefficient, i.e. the dominance coefficient of the most deleterious allele of an individual homozygote for the enhancer locus (i.e. when alleles are

equally expressed). Eq. 2 has been chosen to insure that: (1) the fitness of individual *i* decreases when the most deleterious allele's expression increases, and (2) the more this allele is expressed, the more its effect is revealed.

Additionally to this selection caused by the presence of deleterious mutations, individuals may also undergo stabilizing selection on expression levels. We suppose that there is an optimal expression level Q^*, which corresponds to autosomal expression levels. This dosage requirement results e.g. from the functional coordination of protein activity in pathways involving genes from both autosomes and sex-chromosomes. Expression levels *Q* depend on enhancer log-strengths and expression modifier trait value. We consider three types of expression modifier, corresponding to three classic modes of dosage compensation. The first mode of compensation (Model A) corresponds to an increase in expression of the male X chromosome only (similar to *Drosophila* dosage compensation system). The second mode of compensation (Model B) corresponds to an equal downregulation of the two female X chromosomes (*Caenorhabditis* dosage compensation system). The third mode of compensation (Model C) corresponds to the random extinction of one of the two female X chromosomes (mammal dosage compensation system). Noting m_1 and m_2 the expression modifier trait values associated to the two alleles at the **M** locus, expression levels in the three models (Models A – C) are calculated as:

$$Q_{male} = \left(\frac{m_1 + m_2}{2}\right) Log(e_X) + Log(e_Y) \qquad Q_{fem} = Log\left(e_{X,1}\right) + Log\left(e_{X,2}\right) \tag{3.A}$$

$$Q_{male} = Log(e_X) + Log(e_Y) \qquad Q_{fem} = \left(\frac{m_1 + m_2}{2}\right)\left(Log\left(e_{X,1}\right) + Log\left(e_{X,2}\right)\right) \tag{3.B}$$

$$Q_{male} = Log(e_X) + Log(e_Y) \qquad Q_{fem} = \left(\frac{m_1 + m_2}{2}\right) Log\left(e_{X,1}\right) + Log\left(e_{X,2}\right) \tag{3.C}$$

In model A, the effect of the dosage modifier acts only on the expression of alleles on the X in males. In model B, the effect of the dosage modifier acts only in female and modifies expression levels on both X simultaneously. In model C, the effect of the dosage modifier acts only in female and modifies expression levels on only one of the X. If we note *I* the intensity of stabilizing selection on expression level, then the fitness of individual *i* due to stabilizing selection over expression levels (W_i^E) is:

$$W_i^E = e^{-I(Q_i - Q^*)^2} \quad (4)$$

Total fitness of individual *i* (W_i) is then calculated as $W_i = W^A_i \, W^E_i$. In simulations, two individuals (one male, one female) are drawn randomly from the population, and then accepted as parents of one next-generation individual with a probability equal to their fitness. Then segregation is modelled by randomly choosing one of each parent chromosome to be transmitted to the offspring. In females, the transmission may be preceded by a recombination event between the three loci

All loci undergo recurrent mutations. At each generation, the number of mutations on each locus is drawn from Poisson distributions with mean $N_{pop}\, u_A$ for the gene locus, $N_{pop}\, u_E$ for the enhancer locus and $N_{pop} u_M$ for the modifier locus. u_A, u_E and u_M are the mutation rates per individual per generation for the gene, enhancer and modifier loci, respectively. Mutations occur on chromosomes chosen randomly from the whole population. Mutations on the gene change the fitness effect of the alleles. The mutant fitness effect is drawn from a negative exponential distribution with mean *1/s* (i.e. we only consider deleterious mutations). Mutant enhancer log-strength is drawn from a normal distribution with mean equal to the before-mutation enhancer strength and standard deviation equal to σ_E. On the enhancer locus, mutations impact the log of the strength and not the strength itself to: (1) insure that enhancer strength remains always positive, and (2) prevent mutational variance on relative expression to vanish on ever-increasing mean enhancer strength. Mutant trait values at the expression modifier locus **M** are similarly drawn from a normal distribution with mean equal to the before-mutation trait value, and standard deviation equal to σ_M.

Populations are initialized with no polymorphism: no deleterious alleles on the gene ($W^A_i = 1$ for all individuals), strength of every enhancer copy is equal to *10* (log-strength equal to *1*, for total expression levels equal to $Q = Q^* = 2$) and modifier trait values are all equal to *1*. Then, we perform 1000 generations with mutations occurring only on the gene locus. After 1000 generations, the gene locus is close to mutation-selection equilibrium. Then, 500 000 generations are run with gene, enhancer and modifier mutations and mean trait value of all three loci are recorded at regular time intervals. Simulations are run 100 times, and results are averaged over these replicates. For Fig 5, in each iteration, we calculated at generation

50 000 the average dominance coefficient of alleles at the **A** locus (the gene) carried on Y chromosomes.

Parameter values are set at: mutation rates $u_A = u_E = u_M = 0.0005$, gene purifying selection average intensity $s = 0.1$, standard deviation of mutation normal distribution $\sigma_E = \sigma_M = 0.2$, default dominance coefficient $h = 0.25$, recombination rate between the modifier locus and the enhancer locus $R_{ME} = 0.5$ (only concerns females). Recombination rates between the enhancer locus and the gene locus $R_{EA} = 0.0001$ (only concerns females) unless it is said otherwise, and population size $N_{pop} = 1\ 000$ unless it is said otherwise.

Results

Deleterious mutation accumulation in the absence of cis-regulatory polymorphism

First, we ran simulations without considering regulatory variation. In this model, there is only one gene locus, which undergoes recurrent deleterious mutations. On Fig 1, the mean fitness effect of gene alleles on X and Y chromosomes is illustrated. The results show an increase of this mean fitness effect on X and Y chromosomes. This increase is bigger at lower population sizes, and for Y chromosomes. This shows that, in these simulations, the accumulation of deleterious mutations is due to genetic drift. As there is three times less Y chromosomes than X chromosomes, and as they do not recombine, the effective population size of Y chromosomes is lower than of X chromosomes. Stochastic accumulation of deleterious mutations is thus stronger for Y chromosomes.

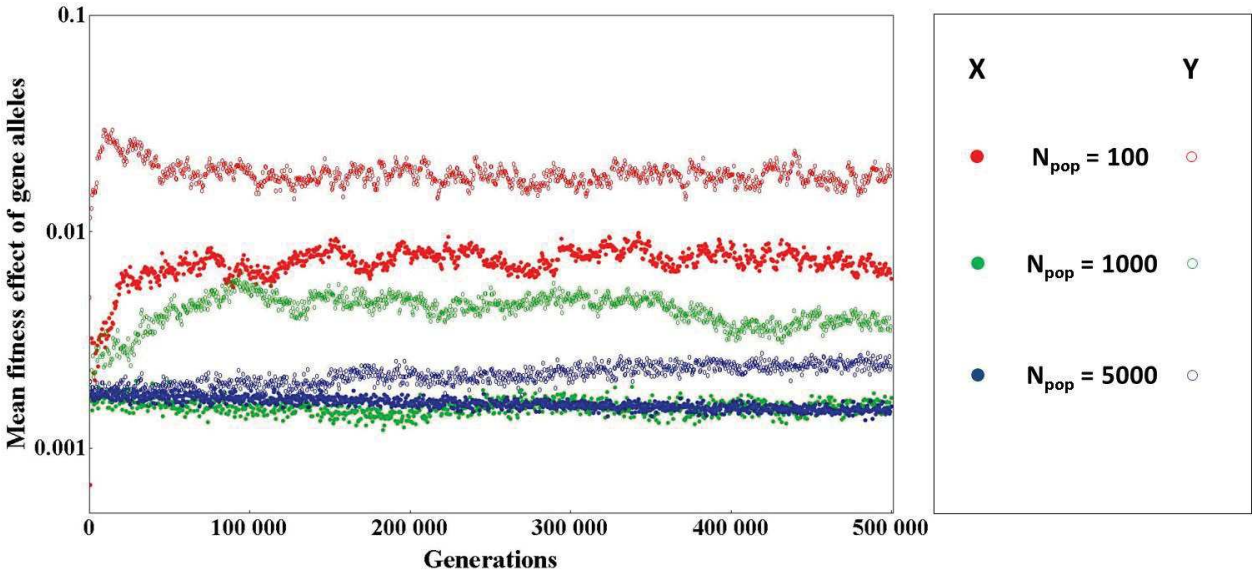


Figure 1. Mean fitness effect of gene alleles located on X (circles) and Y (rings) sex chromosomes. In these simulations, there is no regulatory polymorphism ($u_E = u_M = 0$). Population size is set at $N_{pop} = 100$ (red), $N_{pop} = 1000$ (green) and $N_{pop} = 5000$ (blue). Mean fitness effect of gene alleles is bigger for Y chromosomes than for X chromosomes, and is bigger at low population sizes.

Introduction of cis-regulatory polymorphism

We then introduced *cis*-regulatory polymorphism: a *cis*-regulatory sequence (enhancer) undergoes recurrent mutations changing chromosome-specific expression levels. When a mutation increases the strength of the regulatory sequence, the gene on the same chromosome (and not the homologous chromosome) produces more proteins. Here, we do not consider stabilizing selection on total expression levels (this is done below) to investigate the consequence of regulatory polymorphism without its interaction with stabilizing selection on expression levels. This is like assuming there is a tight negative feedback regulatory loop strictly constraining expression levels at the optimum, irrespectively of the enhancer strength. In this case, *cis*-regulatory polymorphism does not influence total expression levels, but the relative shares of proteins produced from each homologous chromosome. This has a direct selective consequence by changing dominance coefficients of deleterious gene mutations (FYON *et al.* 2015).

On Fig 2, we look at the mean fitness effects of gene alleles on X and Y chromosomes, in cases of *cis*-regulatory polymorphism, for different recombination rates in females between the gene and the enhancer loci. Results presented in Fig 1 for $N_{pop} = 1\,000$ are again illustrated in Fig 2, as a comparison. This comparison clearly shows that mean fitness effect of gene alleles on the Y chromosome increase when there is *cis*-regulatory polymorphism: gene alleles on Y chromosome accumulate more deleterious mutations when there is *cis*-regulatory polymorphism. Y genes tend towards a mean fitness effect of 0.1. This is due to our mutation model, which assumes that deleterious effects of mutations on the gene are drawn from a negative exponential distribution with mean 0.1. Mean deleterious effects of mutation accumulating simply tend to this value. Without such an assumption, for example by allowing several deleterious mutations on the gene to add up, one would expect further increase of mean deleterious effects of gene alleles on Y chromosomes. This faster accumulation of deleterious alleles on the gene in cases of *cis*-regulatory polymorphism indicates that another process is participating to the degeneration of the Y chromosome. This degeneration does not depend on the recombination rate between the gene and enhancer in females.

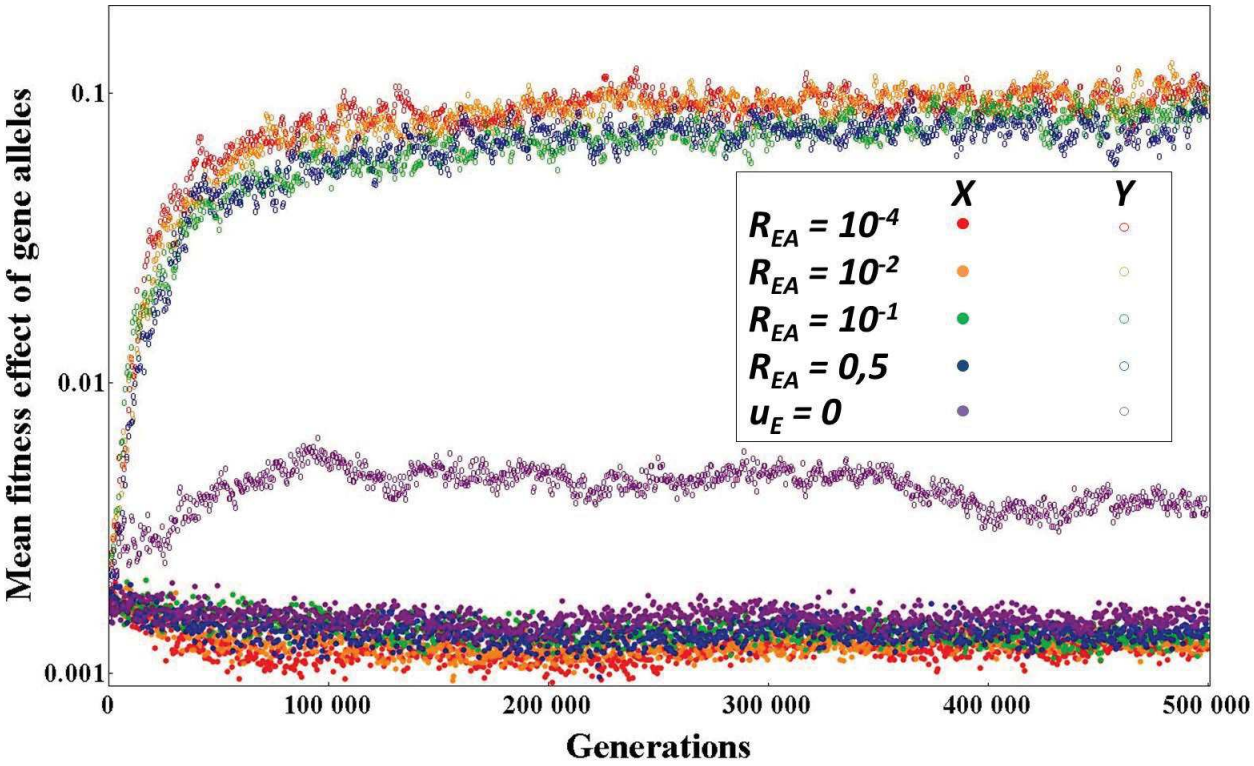


Figure 2. Mean fitness effect of gene alleles on X and Y chromosomes depending on the recombination rate between the gene and enhancer loci in females. Recombination rate values used here are $R_{EA} = 10^{-4}$ (red), $R_{EA} = 10^{-2}$ (orange), $R_{EA} = 10^{-1}$ (green), $R_{EA} = 0.5$ (blue). In purple, we represent mean fitness effect of gene alleles when there is no *cis*-regulatory polymorphism. Circles correspond to X chromosomes, while rings represent Y chromosomes. Mean fitness effect of gene alleles increase on Y chromosomes, and decrease on X chromosomes. Population size is set to *1 000* individuals.

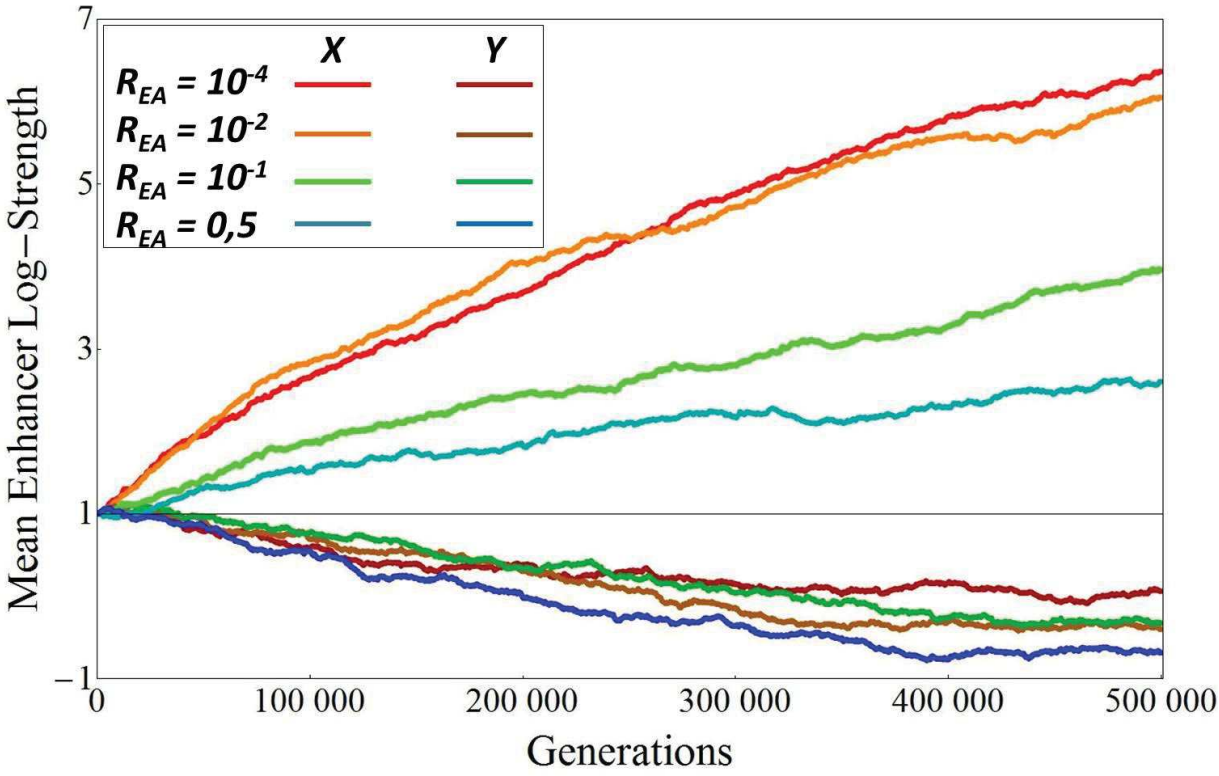


Figure 3. Mean enhancer log-strength of chromosomes X (pale shades) and Y (darker shades) for different recombination rates between the enhancer and gene loci in females. Recombination rate values used are $R_{EA} = 10^{-4}$ (red), $R_{EA} = 10^{-2}$ (orange), $R_{EA} = 10^{-1}$ (green) and $R_{EA} = 0.5$ (blue). Sex chromosomes diverge: X chromosome mean enhancer strength increases, while Y chromosome mean enhancer strength decreases.

To understand what is happening on the enhancer locus that accelerates Y chromosome degeneration, we illustrate on Fig 3 mean enhancer strength on X and Y chromosomes for various recombination rates in females. Fig 3 shows that *cis*-regulatory sequences diverge. On Y chromosomes, enhancer strength tends to decrease while, on X chromosomes, enhancer strength tends to increase. This divergence leads to Y chromosomes being shut down, hidden to selection.

However, looking simultaneously at Fig 2 and 3, deleterious mutations seem to accumulate before chromosomes significantly diverge (i.e. much before generation 100 000 in Fig 2, while divergence in expression seems to be small). This is actually a visual artefact, and mutations on the Y chromosome are already quite hidden. This is because enhancer strength is indicated in log-scale. A difference in one order of magnitude ($log(e_2)$ = 2 compared to

log(e_1) = 1) in enhancer strength causes already a very strong recessivity (*e_1/(e_1+e_2)* small). Also, due to the shape of the relationship between dominance and enhancer strengths, mean dominance of Y chromosome alleles is actually lower than what could be expected from mean enhancer strengths ($\bar{h} < h(\bar{e_1}, \bar{e_2})$).

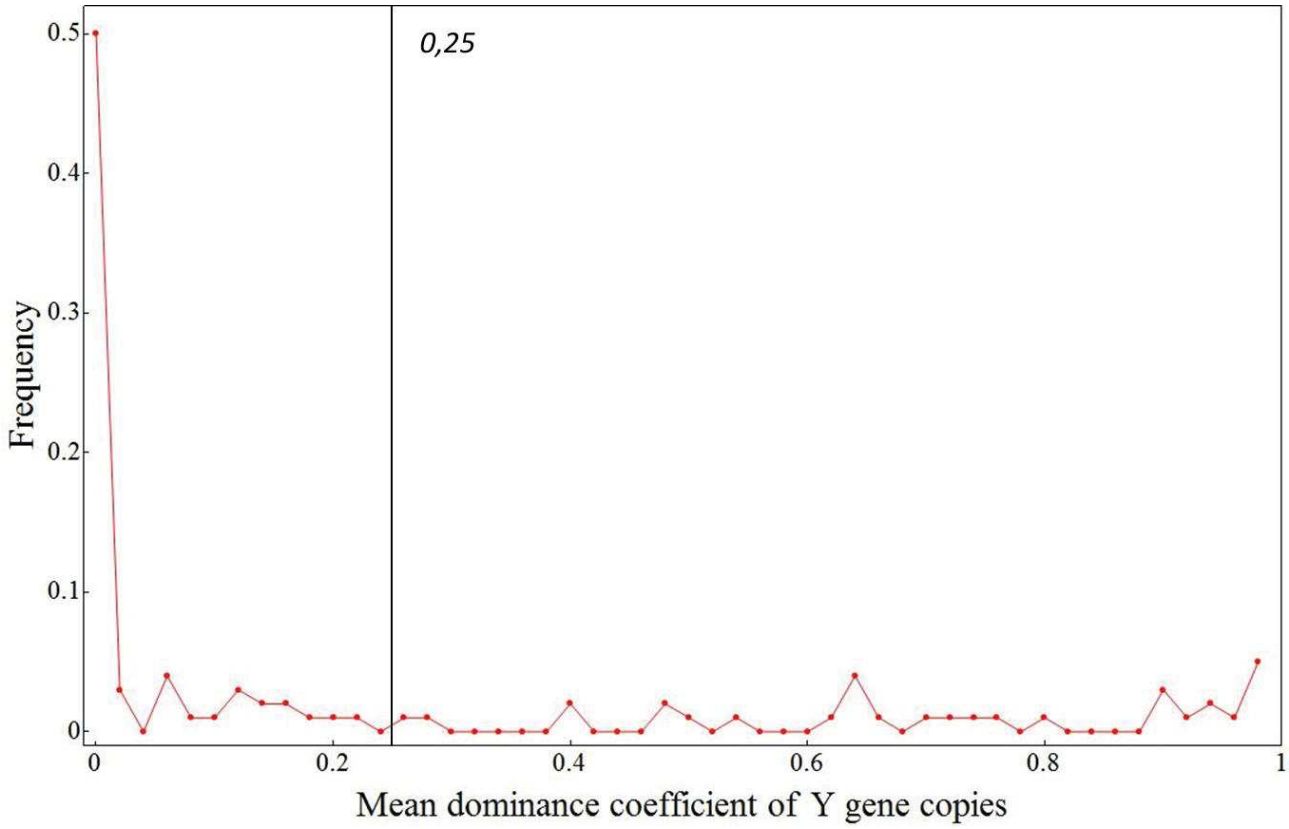


Figure 4. Distribution of mean dominance of Y gene copies after *50 000* generations. Here, our simulations are run 100 times (iterations). For each, we collect, at generation *50 000*, the mean dominance coefficient of gene alleles located on Y chromosomes. Point with an x-axis value of 0.5, for example, represents the frequency of Y gene copy dominance coefficient being equal to a value superior to 0.5 and inferior to 0.52. Mean dominance coefficient is under 0.25 if Y chromosomes are preferentially associated with weaker enhancers, and is over 0.25 if they are associated with stronger enhancers. *R_{EA} = 0,5.*

On Fig 4, we look at the distribution of Y chromosome gene allele dominance coefficients at generation 50 000 estimated using many replicated simulations. Even though chromosome divergence at this generation on Fig 3 looks modest, it corresponds in fact to Y chromosomes being already strongly hidden in most cases (Y dominance coefficient is under 0.1 in 57% of cases). Genes on the Y chromosome are thus quickly hidden from selection (*h* under *0.25*), and accumulate deleterious mutations as weaker enhancers accumulate on the Y and stronger enhancers on the X.

Why do chromosomes diverge? Because of the mode of transmission of sex chromosomes. Indeed, male chromosomes are inherited in a clonal way: male chromosomes do not recombine with the X. As for clonal lineages (see Chapter III), this isolation means that beneficial genetic associations between stronger enhancers and more viable gene alleles accumulate on one chromosome, while associations between weaker enhancers and less viable gene alleles accumulate on the other one.

In clonal lineages, from an initial, non-diverged state, any chromosome can become the stronger-enhancer bearing chromosome, depending on the first mutations and genetic background. For sex chromosomes, however, Y chromosome is always the chromosome accumulating associations between weaker enhancers and less viable gene alleles. This is indeed the only option, as the X cannot be silenced since it is the only chromosome present in females.

Stabilizing selection on expression levels and dosage compensation

We now consider models with stabilizing selection on expression levels in males and females. This selection maintains expression levels around an optimum value, such that an increase of enhancer strength can only evolve if it is compensated by another regulatory change.

Model A

In Model A, male X chromosome expression levels may increase either through *cis*-regulatory changes that also impact female X chromosome expression levels (at locus **E**), or through *trans*-regulatory changes that only impact male X chromosome expression levels (at locus **M**).

We present on Fig 5 mean expression levels of male and female sex chromosomes, and the mean fitness effect of alleles at locus **A** (the gene) depending on the intensity of stabilizing selection on expression levels. Because alleles at locus M do not alter female X chromosome expression levels, expression levels of female X chromosomes are equal to X chromosome mean enhancer strength. They increase when stabilizing selection intensity is low enough to allow for the evolution of sub-optimal expression levels: X chromosome expression levels stabilize to an equilibrium value that balances Enhancer Divergence (ED) process in males and stabilizing selection. Due to the ED process, Y chromosome expression levels (and mean Y enhancer strength) decrease. Y chromosome expression levels decrease more at low stabilizing selection intensity. When $I = 10^{-4}$, male X chromosome expression levels are similar to female X chromosome expression levels: X chromosome expression levels increase mostly through *cis*-regulatory changes (locus **E**). Stabilizing selection intensity allows enough departure from optimal dosage for females to exhibit sex chromosome over-expression. As *I* increases, male X chromosome expression levels increase less, but much more than female X chromosome expression levels, which are limited due to more intense stabilizing selection. As *I* increases, male X chromosome expression level is more influenced by *trans*-regulatory changes (locus **M**) than *cis*-regulatory changes (locus **E**).

Sex chromosome divergence occurs for any intensity of stabilizing selection, but is slower when stabilizing selection is stronger. As previously, we see on Fig 5.e that sex chromosome divergence leads to a fast accumulation of deleterious alleles on Y chromosomes. This accumulation is faster at lower stabilizing selection intensities. In all cases, it is significantly faster than the accumulation expected from the reduction of Y chromosome effective population size.

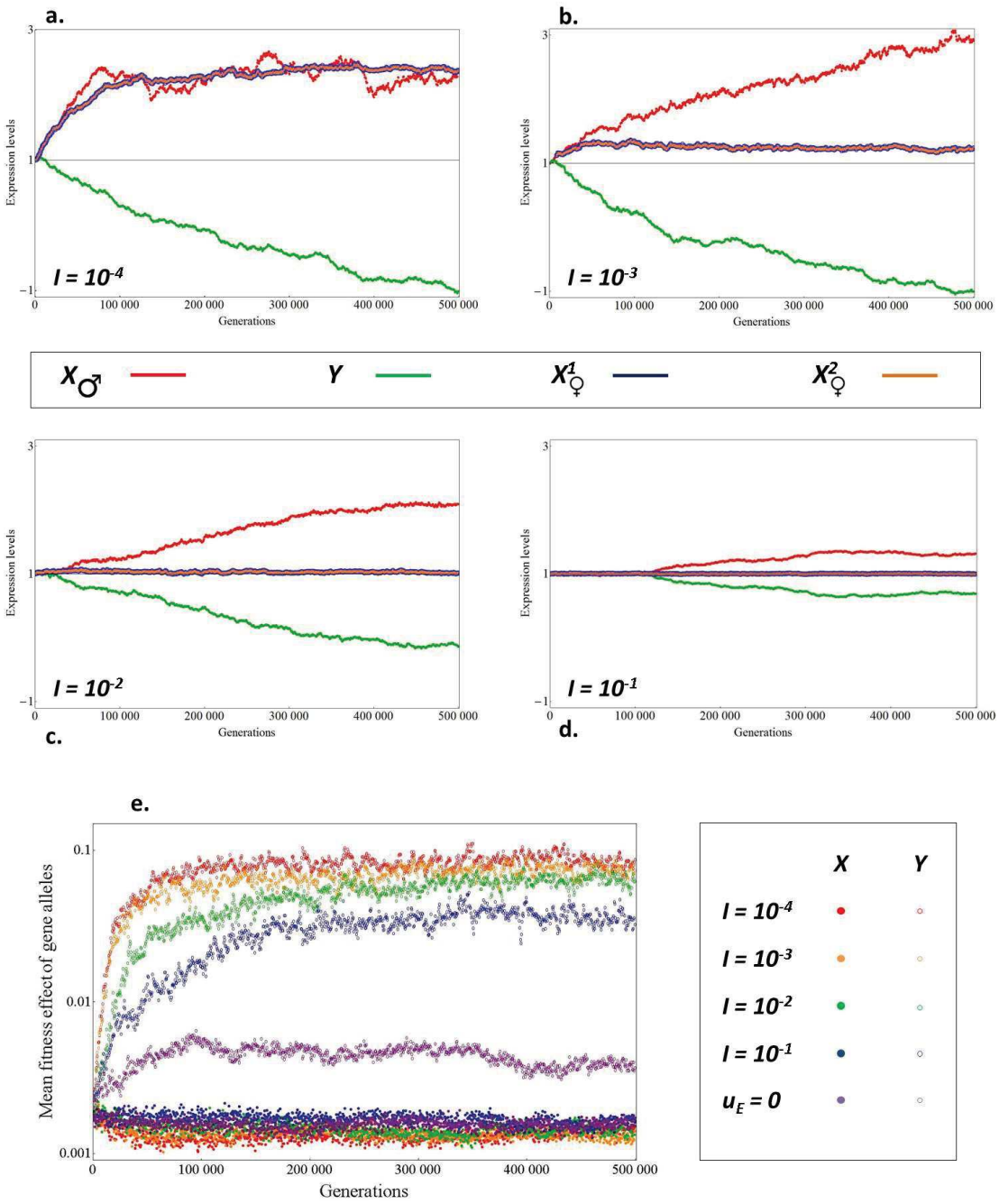


Figure 5. Sex chromosome divergence in Model A for various intensities of stabilizing selection over expression levels. **a – d** Expression levels of male X chromosomes (red), Y chromosomes (green), and female X chromosomes (blue and orange). Different values of stabilizing selection intensity are illustrated: $I = 10^{-4}$ (**a**), $I = 10^{-3}$ (**b**), $I = 10^{-2}$ (**c**), $I = 10^{-1}$ (**d**). **e** Mean fitness effect of gene alleles on X (circles) and Y (rings) chromosomes. Stabilizing selection intensities used here are the same: $I = 10^{-4}$ (red), $I = 10^{-3}$ (orange), $I = 10^{-2}$ (green), $I = 10^{-1}$ (blue). In purple, we show mean fitness effect of gene alleles in a case of no *cis*-regulatory polymorphism, as a comparison.

Model B

In Model B, male X chromosome expression levels increase only through *cis*-regulatory changes that also impact female X chromosome expression levels. However, female sex chromosome expression levels can be restored to optimal dosage through *trans*-regulatory changes (at locus **M**) that impact expression of both females X chromosomes equally.

Results are presented on Fig 6. As previously, we see that male sex chromosomes diverge: X chromosome mean enhancer strength (and expression levels) increases and Y chromosome mean enhancer strength (and expression levels) decreases. Female X chromosome expression levels do not increase and are maintained around optimal dosage. *Trans*-regulatory changes in females evolve to compensate for the increased strength of enhancers on the X.

Again, sex chromosome divergence occurs at all stabilizing selection intensities used, although it is slower for stronger stabilizing selection. The accumulation of deleterious mutation on Y chromosomes is similar to the one observed in Model A: faster at lower selection stability intensities, and always faster than the accumulation occurring in absence of regulatory changes.

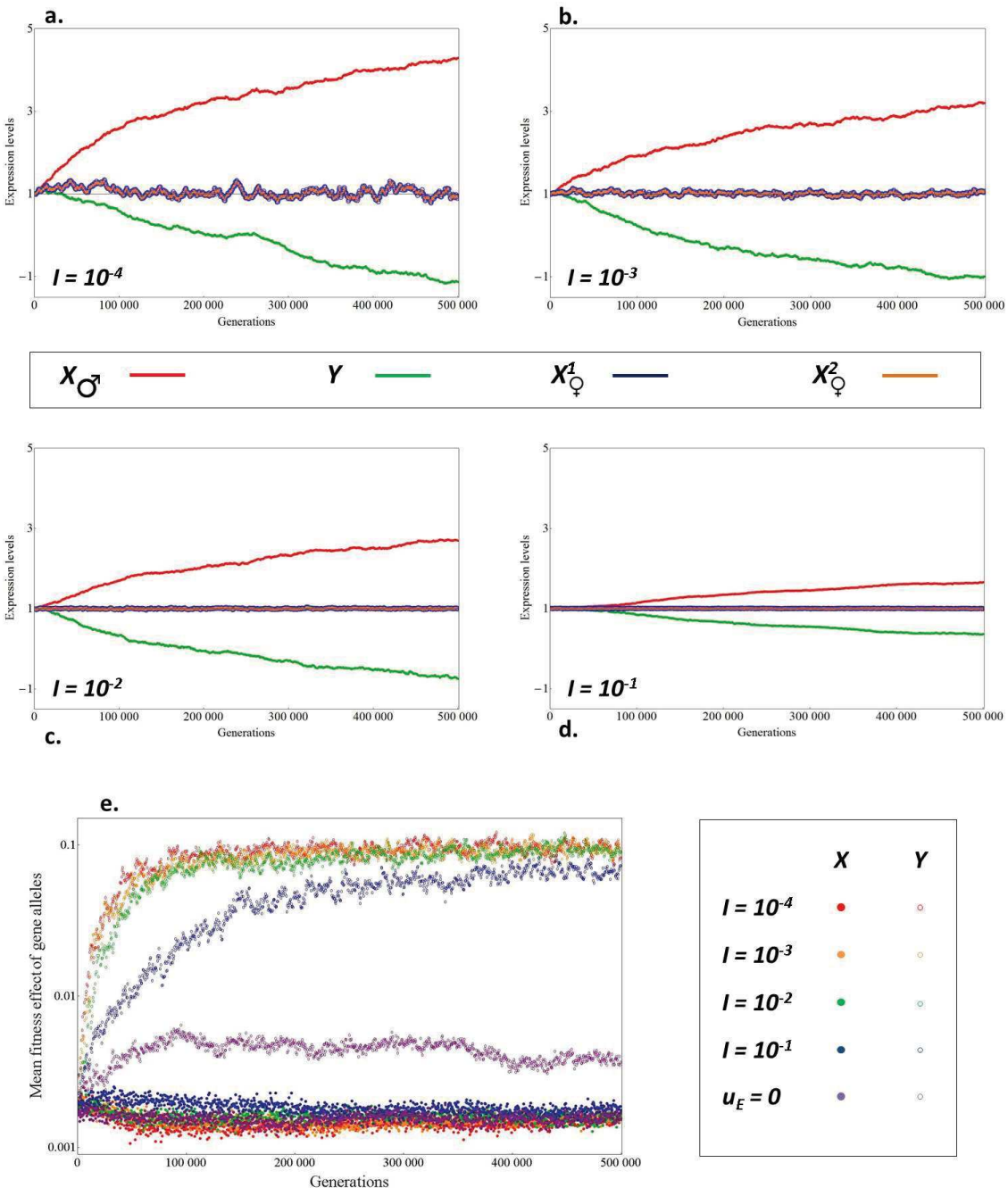


Figure 6. Sex chromosome divergence in Model B for various intensities of stabilizing selection over expression levels. **a – d** Expression levels of male X chromosomes (red), Y chromosomes (green), and female X chromosomes (blue and orange). We use various values of stabilizing selection intensity: $I = 10^{-4}$ (**a**), $I = 10^{-3}$ (**b**), $I = 10^{-2}$ (**c**), $I = 10^{-1}$ (**d**). **e** Mean fitness effect of gene alleles on X (circles) and Y (rings) chromosomes. Stabilizing selection intensities used here are the same: $I = 10^{-4}$ (red), $I = 10^{-3}$ (orange), $I = 10^{-2}$ (green), $I = 10^{-1}$ (blue). In purple, we show mean fitness effect of gene alleles in a case of no *cis*-regulatory polymorphism, as a comparison.

Model C

Model C is similar to Model B, except that *trans*-regulatory changes at locus **M** impact only one female X chromosome expression levels. Results are presented on Fig 7.

As before, male sex chromosomes diverge: X chromosome mean enhancer strength (and expression levels) increases as much as Y chromosome mean enhancer strength (and expression levels) decreases. In females, one X chromosome expression levels increase as much as male X chromosome expression levels: this is the unmodified female X chromosome. Meanwhile, the other female X chromosome expression levels decrease through *trans*-regulatory changes (locus **M**). When $I = 10^{-4}$, they do not decrease as much as Y chromosome expression levels, as stabilizing selection intensity is low enough to allow for female over-expression. At higher *I* values, the decrease of the expression level of the silenced X is comparable to the silencing level observed on the Y.

Sex chromosome divergence occurs again at all intensities of stabilizing selection, and is faster at low stabilizing selection intensities. Deleterious mutation accumulation is similar to Model A and B.

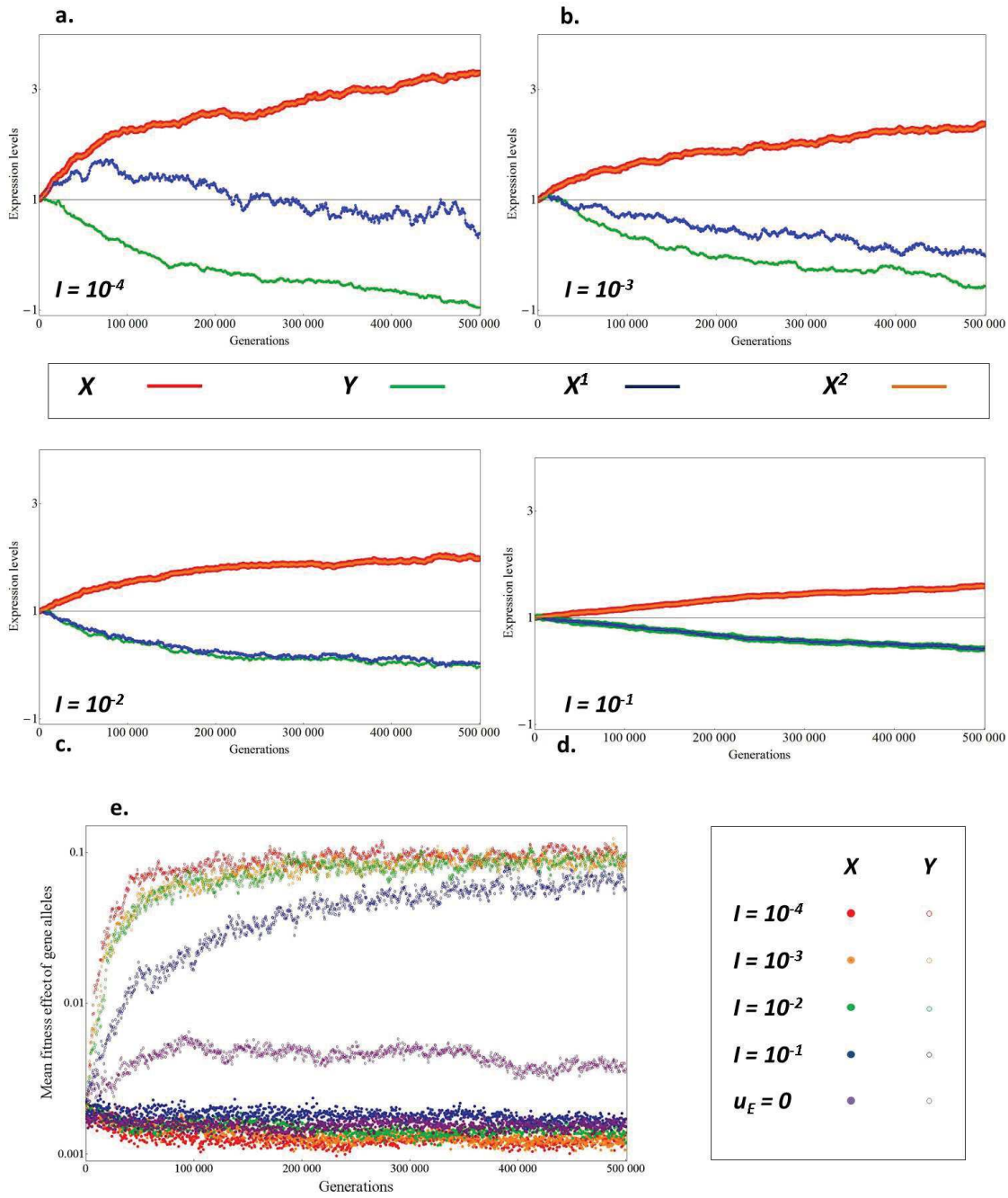


Figure 7. Sex chromosome divergence in Model C for various intensities of stabilizing selection over expression levels. **a – d** Expression levels of male X chromosomes (red), Y chromosomes (green), and female X chromosomes (blue and orange). We use various values of stabilizing selection intensity: $I = 10^{-4}$ (**a**), $I = 10^{-3}$ (**b**), $I = 10^{-2}$ (**c**), $I = 10^{-1}$ (**d**). **e** Mean fitness effect of gene alleles on X (circles) and Y (rings) chromosomes. Stabilizing selection intensities used here are the same: $I = 10^{-4}$ (red), $I = 10^{-3}$ (orange), $I = 10^{-2}$ (green), $I = 10^{-1}$ (blue). In purple, we show mean fitness effect of gene alleles in a case of no *cis*-regulatory polymorphism, as a comparison.

Discussion

Originality of the model

Previous models have tried to explain Y chromosome degeneration through selective interference associated with the suppression of recombination on the Y chromosome (CHARLESWORTH 1996). In a nutshell, these models showed that, on non-recombining chromosomes, deleterious mutations are more likely to be fixed, while beneficial mutations are less likely to be fixed. These models were investigated assuming multiple non-epistatic loci under selection. This process has been investigated with little genetic details. For instance, Muller's ratchet on the Y have been investigated with haploid models (see, for instance, GORDO and CHARLESWORTH 2001). Some models have formally considered sexual diploid populations and sex chromosomes (CHARLESWORTH and CHARLESWORTH 1997; ENGELSTÄDTER 2008), but they still focus on Y chromosome accumulating fitness lag due to selective interference. They do not study Y chromosome expression extinction, and thus do not model gene expression regulation: they only assume that Y chromosome fitness lag will lead to an adaptive extinction of Y chromosome expression. Yet, gene expression regulation is a major target of evolution (WRAY *et al.* 2003). Moreover, Y chromosome extinction and dosage compensation on X chromosomes (which aims at balancing expression levels between autosome pairs and sex chromosome pairs) can only occur through changes in gene expression.

Here, we used radically different models to uncover another process that may be a significant driver of sex chromosome evolution. First, our individual-based models specifically consider sex chromosomes. Second, they formally incorporate gene expression regulation, through an enhancer locus that *cis*-regulates a focal gene, and through a *trans*-regulatory locus that alters sex chromosome expression levels. Importantly, as in previous models, we do not consider sex-determining genes, as they are expected to evolve very differently from other sex chromosome genes. Male-specific genes, notably, are not expected to undergo the genetic degeneration that characterizes other Y chromosome genes.

Accumulation of deleterious mutations on Y chromosomes

In classic theory of sex chromosome evolution, it is thought that Y chromosome fitness lags behind X chromosome fitness because of selection interferences between non-recombining loci. This can be interpreted as a reduction of effective population size, as selection on each locus is less efficient due to selection at linked loci. Non-recombining Y chromosomes have a low effective population size, such that deleterious mutations are more likely to be fixed than in absence of interference.

Our results show that another process may accelerate the spread of deleterious mutations on Y chromosomes. Non-recombining X and Y chromosomes in males are prone to enhancer divergence, as the absence of recombination leads to the beneficial associations of stronger enhancers with more viable gene alleles, and of weaker enhancers with less viable gene alleles. In the general case, we have seen in Chapter I that these associations lead to the invasion of the stronger enhancers, as they benefit from a good genetic background. In the particular case of genetically isolated homologs, we saw in Chapter III that this invasion cannot happen. Instead, homologous chromosomes diverge: in every clonal lineage, one copy of every gene gets purged from deleterious mutations and is associated with ever-stronger enhancers, while the other copy accumulate deleterious mutations and is associated with ever-weaker enhancers. Here, we saw that this divergence process also concerns sex chromosomes, since X and Y chromosomes are genetically isolated. Interestingly, the divergence process for sex chromosomes is asymmetrical, as weaker enhancers always accumulate on the Y. In a clonal lineage where silencing is random between the two homologs (see Chapter III), all individuals in the next generations still inherit a non-silenced copy of the gene. With sex-chromosomes, this is not the case: if the X is silenced, expression is completely shut down in all females. If the Y is silenced, on the contrary, all individuals (males and females) still have at least one active copy of the gene. Hence, enhancer divergence, caused by recombination suppression between X and Y, leads to the asymmetrical silencing of the Y.

Y downregulation and X upregulation

Y downregulation is usually interpreted as a compensatory mechanism to silence deleterious mutations accumulating on Y chromosomes (ORR and KIM 1998). X upregulation is thought to be a dosage compensation mechanism, to ensure that dosage-sensitive genes on sex chromosome are still produced in fitted amountscompared to autosome genes (ERCAN 2015).

In the process studied here, the accumulation of deleterious alleles, Y chromosome downregulation and X upregulation are simultaneous. As X and Y enhancers diverge, it becomes easier for deleterious gene alleles on the Y to fix. As deleterious mutations accumulate, it becomes easier for enhancer-strength decreasing mutations to fix on the Y, and for enhancer strength-increasing mutations to fix on the X.

Dosage compensation

Following Ohno's initial theory on the 'peril of hemizygosity' (OHNO 1967; HALL and WAYNE 2013), modifications of expression levels of X chromosomes have been first interpreted as ways to maintain optimal dosage with autosome expression levels while Y expression is progressively shut down (DISTECHE 2012; ERCAN 2015). For example, in mammals, it is thought that expression levels of dosage-sensitive X genes have been upregulated (NGUYEN and DISTECHE 2006; DENG *et al.* 2011), to compensate for the halving of sex chromosome expression in males. Later, inactivation of one female X chromosomes (MOREY and AVNER 2011) would have evolved to compensate for the doubling of dosage in females. Similarly, in *Caenorhabditis*, X chromosome dosage-sensitive genes have been up-regulated (DENG *et al.* 2011) and then a system has evolved to downregulate X chromosomes of XX hermaphrodites (MEYER 2010). It is only in flies that such a two-step compensation system would not have evolved: a simple up-regulation of male X dosage-sensitive genes would have restored optimal dosage in males (CONRAD and AKHTAR 2012), while dosage was never disturbed in females.

It has been pinpointed that the number of dosage-sensitive genes is probably low, such that all expression modifications may not be explained by the need to produce fitted dosage between autosomes and sex chromosomes (PESSIA *et al.* 2014).

In this paper, we propose another theory to account for the evolution of sex chromosome expression levels. We argue that X chromosome upregulation does not evolve for maintaining proper dosage with autosomes in males, but rather as a direct product of chromosome divergence in males. Indeed, we found that X chromosome enhancer strength increases even when we considered no stabilizing selection on expression levels (so that there is no need for dosage compensation). Two possibilities for X chromosome upregulation arise. First, up-regulation occurs through *cis*-regulatory changes. Then, X upregulation also occurs in females, which starts producing too many proteins. In dosage-sensitive genes, one female X chromosome is progressively shut down (mammals), or both female X chromosomes get downregulated (*Caenorhabditis*), through *trans*-regulatory changes. Alternatively, male X upregulation occurs through changes that apply only to males (flies). Then, there is no expression level change in females, and thus no dosage compensation is needed.

Dosage compensation mechanisms, until now, invoked either a two-step mechanism for mammals and worms, or a one-step mechanism for flies. Here, we argue that the first step in mammals and worms, and the only step for flies, are actually not part of the dosage compensation mechanism, but rather is the result of chromosome divergence due to indirect selection on regulatory sequences. In systems like mammals and worms, X upregulation during chromosome divergence also occurs in females, requiring dosage compensation changes for dosage-sensitive genes in females, while this is not the case for systems like fly.

Speed of Y degeneration

Our simulations show that deleterious allele accumulation on Y chromosomes is more important when explicitly considering regulatory changes (Fig 2) than when only considering the lower effective population size of Y chromosomes (Fig 1). This means that the Enhancer Divergence process is susceptible to be an important force driving sex chromosome evolution. However, indirect selection on regulatory sequences is weak: of the order of magnitude of gene mutation rate. It is thus probable than ED process is weak for small populations. For small populations, selective interference is probably a more important factor, although the two mechanisms have yet to be quantitatively compared. For large or

very large populations, the stochastic fixation of deleterious mutations due to selective interference is quantitatively less important. In those cases, the process described in this paper may play a stronger, and perhaps predominant role in the evolution of sex chromosomes.

References

BACHTROG D., 2006a Expression Profile of a Degenerating Neo-Y Chromosome in Drosophila. Curr. Biol. **16**: 1694–1699.

BACHTROG D., 2006b A dynamic view of sex chromosome evolution. Curr. Opin. Genet. Dev. **16**: 578–585.

BIRKY C. W., WALSH J. B., 1988 Effects of linkage on rates of molecular evolution. Proc. Natl. Acad. Sci. **85** : 6414–6418.

BULL J. J., 1983 *Evolution of sex determining mechanisms*. Benjamin/Cummings Pub. Co.

CHARLESWORTH B., 1978 Model for evolution of Y chromosomes and dosage compensation. Proc. Natl. Acad. Sci. U. S. A. **75**: 5618–5622.

CHARLESWORTH B., 1994 The effect of background selection against deleterious mutations on weakly selected, linked variants. Genet. Res. **63**: 213–227.

CHARLESWORTH B., 1996 The evolution of chromosomal sex determination and dosage compensation. Curr. Biol. **6**: 149–162.

CHARLESWORTH D., CHARLESWORTH B., 1980 Sex differences in fitness and selection for centric fusions between sex-chromosomes and autosomes. Genet. Res. (Camb). **35**: 205–214.

CHARLESWORTH B., CHARLESWORTH D., 1997 Rapid fixation of deleterious alleles can be caused by Muller's ratchet. Genet. Res. **70**: 63–73.

CHARLESWORTH B., CHARLESWORTH D., 2000 The degeneration of Y chromosomes. Philos. Trans. R. Soc. Lond. B. Biol. Sci. **355**: 1563–1572.

CONRAD T., AKHTAR A., 2012 Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning of chromosome-wide transcription. Nat Rev Genet **13**: 123–134.

DENG X., HIATT J. B., NGUYEN D. K., ERCAN S., STURGILL D., HILLIER L. W., SCHLESINGER F., DAVIS C. A., REINKE V. J., GINGERAS T. R., SHENDURE J., WATERSTON R. H., OLIVER B., LIEB J. D., DISTECHE C. M., 2011 Evidence for compensatory upregulation of expressed X-linked genes in mammals, Caenorhabditis elegans and Drosophila melanogaster. Nat Genet **43**: 1179–1185.

DISTECHE C. M., 2012 Dosage compensation of the sex chromosomes. Annu. Rev. Genet. **46**:

537–560.

ELLEGREN H., 2011 Sex-chromosome evolution: recent progress and the influence of male and female heterogamety. Nat Rev Genet **12**: 157–166.

ENGELSTÄDTER J., 2008 Muller's Ratchet and the Degeneration of Y Chromosomes: A Simulation Study. Genetics **180**: 957–967.

ERCAN S., 2015 Mechanisms of x chromosome dosage compensation. J. genomics **3**: 1–19.

FELSENSTEIN J., 1974 The Evolutionary Advantage of Recombination. Genetics **78**: 737–756.

FYON F., CAILLEAU A., LENORMAND T., 2015 Enhancer Runaway and the Evolution of Diploid Gene Expression. PLoS Genet **11**: e1005665.

GORDO I., CHARLESWORTH B., 2000 On the Speed of Muller's Ratchet. Genetics **156**: 2137–2140.

GORDO I., CHARLESWORTH B., 2001 The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes. Genet. Res. **78**: 149–161.

GRIBNAU J., GROOTEGOED J. A., 2012 Origin and evolution of X chromosome inactivation. Curr. Opin. Cell Biol. **24**: 397–404.

HALL D. W., WAYNE M. L., 2013 Ohno's "Peril of Hemizygosity" Revisited: Gene Loss, Dosage Compensation, and Mutation. Genome Biol. Evol. **5** : 1–15.

HILL W. G., ROBERTSON A., 1966 The effect of linkage on limits to artificial selection. Genet. Res. (Camb). **8**: 269–294.

IRONSIDE J. E., 2010 No amicable divorce? Challenging the notion that sexual antagonism drives sex chromosome evolution. BioEssays **32**: 718–726.

JABLONKA E. V. A., LAMB M. J., 1990 The Evolution of Heteromorphic Sex Chromosomes. Biol. Rev. **65**: 249–276.

LENORMAND T., 2003 The evolution of sex dimorphism in recombination. Genetics **163**: 811–822.

MANNING J. T., THOMPSON D. J., Muller's ratchet and the accumulation of favourable mutations. Acta Biotheor. **33**: 219–225.

MAYNARD J., HAIGH J., 2007 The hitch-hiking effect of a favourable gene. Genet. Res. (Camb).

89: 391–403.

MEYER B. J., 2010 Targeting X chromosomes for repression. Curr. Opin. Genet. Dev. **20**: 179–189.

MING R., WANG J., MOORE P. H., PATERSON A. H., 2007 Sex chromosomes in flowering plants. Am. J. Bot. **94** : 141–150.

MOREY C., AVNER P., 2011 The Demoiselle of X-Inactivation: 50 Years Old and As Trendy and Mesmerising As Ever. PLoS Genet **7**: e1002212.

MULLER H. J., 1964 The relation of recombination to mutational advance. Mutat. Res. Mol. Mech. Mutagen. **1**: 2–9.

NGUYEN D. K., DISTECHE C. M., 2006 Dosage compensation of the active X chromosome in mammals. Nat Genet **38**: 47–53.

OHNO S., 1967 *Sex Chromosomes and Sex-Linked Genes*. Springer-Verlaag, Berlin, Heidelberg, New-York.

ORR H. A., KIM Y., 1998 An Adaptive Hypothesis for the Evolution of the Y Chromosome. Genetics **150**: 1693–1698.

PESSIA E., ENGELSTÄDTER J., MARAIS G. A. B., 2014 The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis? Cell. Mol. Life Sci. **71**: 1383–1394.

PESSIA E., MAKINO T., BAILLY-BECHET M., MCLYSAGHT A., MARAIS G. A. B., 2012 Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome. Proc. Natl. Acad. Sci. U. S. A. **109**: 5346–5351.

RICE W. R., 1987a The Accumulation of Sexually Antagonistic Genes as a Selective Agent Promoting the Evolution of Reduced Recombination between Primitive Sex Chromosomes. Evolution (N. Y). **41**: 911–914.

RICE W. R., 1987b Genetic Hitchhiking and the Evolution of Reduced Genetic Activity of the Y Sex Chromosome. Genetics **116**: 161–167.

WRAY G. A., HAHN M. W., ABOUHEIF E., BALHOFF J. P., PIZER M., ROCKMAN M. V, ROMANO L. A., 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–419.

ZHOU Q., BACHTROG D., 2012 Chromosome-Wide Gene Silencing Initiates Y Degeneration in

Drosophila. Curr. Biol. **22**: 522–525.

Jusqu'à présent, nous nous sommes intéressés aux aspects théoriques de la sélection sur les séquences régulatrices. Nous avons élaboré de nombreux modèles afin de voir les conséquences de cette sélection indirecte avec ou sans sélection stabilisante sur les niveaux d'expression, dans le cadre de différents modes de reproduction, et dans le cas de chromosomes sexuels. Ces différents cas de figures nous ont permis de comprendre l'origine de cette sélection indirecte ainsi que les conséquences évolutives qu'elle est susceptible d'entraîner. De façon importante, cela nous a permis d'établir un certain nombre de prédictions.

De nombreuses prédictions de nos modèles sont compatibles avec ce que l'on connaît des réseaux de régulation. Ainsi, on prédit l'établissement de réseaux de régulation inutilement compliqués, du fait du recrutement de nouveaux régulateurs comme façon d'augmenter la force des *cis*-régulateurs (voir Annexe). De même, on prédit que des mutations aléatoires sur les *cis*-régulateurs devraient être biaisées pour diminuer les niveaux d'expression (puisque les *cis*-régulateurs, après de nombreuses générations, ont fini par atteindre une force quasi-maximale, voir Chapitre I), tandis que des mutations aléatoires sur les *trans*-régulateurs devraient être biaisées pour augmenter les niveaux d'expression, ce qui a été déjà été observé (METZGER *et al.* 2016).

Parmi ces prédictions, certaines peuvent faire l'objet de tests empiriques. Ainsi, nos modèles prédisent que les *cis*-régulateurs d'espèces allogames devraient être plus forts que les mêmes *cis*-régulateurs d'espèces autogames proches (Chapitres I et III). Une hybridation double sens (chaque espèce sert tour à tour de père et de mère aux hybrides) permettrait de tester cette prédiction. En effet, des différences d'expression entre homologues chez l'hybride F1 peuvent être uniquement attribuées à des différences de *cis*-régulateurs, puisque les homologues partagent le même contexte *trans*-régulateur. On devrait donc observer, chez l'hybride, une expression plus forte de l'haplotype venant du parent allogame. Toutefois, les rares exemples de tels hybrides dans la littérature n'étaient pas double sens, de sorte que l'on ne peut conclure entre *ER process* et effets parent-d'origine.

Dans le dernier chapitre, nous utilisons des données de divergence entre *Mus musculus* et *Rattus norvegicus* ainsi que des données de transcriptomique pour tester certaines de nos prédictions. Nous définissons un signal qui ne peut être expliqué que par l'*ER process* : les

cis-régulateurs devraient évoluer plus rapidement (et donc diverger plus vite entre espèces) quand ils sont situés près du gène. Ils devraient également diverger davantage que par un processus neutre, puisque l'*ER* process est issu d'une pression de sélection positive. Nos résultats montrent que ce signal est faible, mais détectable, dans les données que nous avons utilisées. Il s'agit là de la première confirmation empirique de la sélection indirecte sur les *cis*-régulateurs que nous avons théorisés. A noter que la faiblesse du signal ne présage pas forcément d'une faiblesse du processus. En effet, les signaux de sélection positive sont noyés dans des signaux de sélection stabilisante très présents dans les réseaux de régulation. Qui plus est, il y a une véritable difficulté à cibler des séquences que l'on sait être des séquences régulatrices. Nous étudions donc des « régions » vraisemblablement enrichies en séquences régulatrices, ce qui dilue le signal.

Paper in preparation

Chapitre V

Finding a genomic footprint specific to the 'Enhancer Runaway' Process

Authors: Frederic Fyon[1], Gustavo Barroso[2], Julien Dutheil[2,3 *], Thomas Lenormand[1 *]

[1] CEFE, UMR 5175, Montpellier, France

[2] Max Planck Institute for Evolutionary Biology, Evolution Department of Evolutionary Genetics, Plön - - GERMANY

[3] Institut des Sciences de l'Evolution, Université de Montpellier, France

* Co-last author

Introduction

Gene expression regulation is a major element of the genotype-to-phenotype map. Evidence accumulates of evolutionary novelties occurring through regulatory changes (RAYMOND *et al.* 1998; FEREA *et al.* 1999; COOPER 2003; ABZHANOV *et al.* 2004; SHAPIRO *et al.* 2004). Deciphering the various evolutionary forces acting upon gene expression regulation is thus of prime importance in understanding species evolution. Many methods have been developed to study the relative influence on expression evolution of stabilizing selection, genetic drift and directional selection (FAY and WITTKOPP 2008). These methods can be classified into two broad categories.

First category includes comparative methods. Several studies have performed comparisons of expression patterns along phylogenies, showing that some of them correlate with environmental conditions like temperature or adaptive traits like copper sulfate resistance (FAY *et al.* 2004). Phylogenetic comparisons may also concern expression disparities among species. For example, contradictory works have been published about accelerated (COOPER 2003; KHAITOVICH *et al.* 2005) or decelerated (HSIEH *et al.* 2003; GILAD *et al.* 2006) rates of expression changes within the human lineage. In comparative studies, phylogenetic relationships need to be taken into account to decipher between historical and adaptive correlations. However, it is more difficult to rule out patterns of changes that are due to genetic and ontogenetic correlations: expression changes may not be due to selection, but associated with another locus whose variation is under selection.

Second category covers neutrality test methods. This approach consists in comparing data to neutral expectation. The latter strongly depends on the mutational pattern and needs to be adapted to the case study. This approach is based on the ratio of expression divergence between species over expression variation within species. Directional selection is inferred when this ratio exceeds its neutral expectation. On the contrary, lower than neutral ratios are a signal of stabilizing selection. Studies using such methods have identified widespread signals of stabilizing selection, and few cases of directional selection (RIFKIN *et al.* 2003; UDDIN *et al.* 2004). A major issue with this approach is that the assumptions made to obtain the neutral expectation are hard to verify. In some studies, neutral expectation is derived from empirical proxies. For example, some studies use pseudogenes to measure neutral variation

within species and neutral divergence between species (KHAITOVICH *et al.* 2004). Others use mutation-accumulation experiments (DENVER *et al.* 2005; RIFKIN *et al.* 2005).

Overall, a global view of expression evolution slowly emerges (WHITEHEAD and CRAWFORD 2006a). Stabilizing selection is the most widespread selective force influencing expression evolution (LUDWIG *et al.* 2000; DENVER *et al.* 2005; GILAD *et al.* 2006). Many studies also documented patterns of regulatory evolution consistent with random genetic drift (OLEKSIAK *et al.* 2002; KHAITOVICH *et al.* 2004; YANAI *et al.* 2004). Neutral evolution may be due to synonymous mutations, or due to the fact that different regulatory networks may lead to the same phenotype (WEIRAUCH and HUGHES 2010). Finally, directional selection seems to be occasionally involved during adaptation events (NUZHDIN *et al.* 2004; WHITEHEAD and CRAWFORD 2006b).

In a recent paper (FYON *et al.* 2015), we described a new type of positive selection pressure acting on regulatory sequences. This type of selection pressure was never mentioned previously and therefore never investigated in the list of evolutionary forces influencing gene expression evolution. This selection is rooted in the phenotypical consequences of *cis*-regulatory polymorphism and allele-specific expression. Indeed, *cis*-regulatory sequences – namely, enhancers and promoters – are defined as regulating the expression of only one gene copy located on the same chromosome. As a consequence, *cis*-regulatory polymorphism induces allele-specific expression, which has been found to be widespread (LO *et al.* 2003). A difference in the ability to promote transcription ('strength') among homologous enhancers is likely to create an unbalance in the amount of proteins that are produced from homologous gene copies. In other words, a stronger enhancer copy will lead to its associated gene copy contributing a larger share of proteins. Because the final phenotype is more influenced by the gene allele associated with the stronger enhancer, this gene allele is more visible to selection. As a result, genes associated to stronger enhancers tend to be more efficiently purged from deleterious mutations. If the recombination rate between the enhancer and gene loci is not too high, a positive genetic association is built between stronger enhancer alleles and more viable genetic background, and between weaker enhancers and less viable genetic background. These associations lead to an indirect positive selection for stronger enhancers close to the genes (proximal enhancers). Due to this selection, proximal enhancer strength escalates in an open-ended fashion until physical

limitation, a process that we referred to as the 'Enhancer Runaway' (ER) process (FYON *et al.* 2015).

Interestingly, the ER process may not be halted by stabilizing selection on expression levels. Indeed, overexpression due to increased enhancer strength may be compensated through mutations on other *cis*- or *trans*-regulatory elements (see Chapters I & II). This would result in regulatory evolution to run idle: regulatory sequences change, whole regulatory networks would change over time, but without significantly altering expression levels. This disconnection between regulatory evolution and phenotypical evolution has already been documented, and is argued to be due to an equivalency among various regulatory networks (WEIRAUCH and HUGHES 2010). As different networks may lead to identical phenotypes, it is argued that evolution between these networks is neutral and affected by stabilizing selection (mutations disrupting expression levels being compensated by other mutations restoring expression levels). In Fyon et al. (2015), we argued that evolution between these equivalent networks is not neutral, but rather tends to favor networks with strong enhancers close to the gene. More importantly, we argued that the ER process works as a driving force accelerating the evolution between equivalent networks (FYON *et al.* 2015).

Here, we seek to provide the first empirical tests of ER process theory. We use measures of transcriptional noise, obtained from single-cell transcriptomic data of the domestic mouse *Mus musculus*, as a proxy for the presence of negative feedback regulatory loops. We combine these data with measures of regulatory and non-regulatory sequence divergence between the genomes of *Mus musculus* and *Rattus norvegicus* in order to detect specific footprint of the ER process. We indeed expect that this process should proceed faster than neutral, and faster for regulatory sequences closer from genes, impacting divergence patterns accordingly.

The genomic footprint of the ER process

As explained above and in the first paper originally describing the ER process (FYON *et al.* 2015), competition for expression between *cis*-regulatory sequences creates an indirect selection favoring stronger *cis*-regulators close to the gene. Overall, this accelerates *cis*-regulatory sequence turn-over: new stronger *cis*-regulators are continually being favored. This also promotes coevolution between regulatory sequences, which allows compensating for the escalation of the strength of *cis*-regulatory elements located at short distance from genes. This compensation may involve both distant *cis*-regulatory elements and *trans*-regulatory elements (Chapter II). The ER process thus promotes the turn-over of all regulatory sequences. As a consequence, it is expected to have significant impact on regulatory divergence patterns between closely-related species. More precisely, we expect that a stronger ER process should lead to faster divergence of regulatory sequences. Regulatory divergence between closely-related species is usually interpreted as the result of neutral or adaptive processes. Stabilizing and background selection, on the contrary, tend to limit regulatory divergence between closely-related species. Here, we aim to find patterns among regulatory sequence divergence between *Mus musculus* and *Rattus norvegicus* that leave a footprint which is specific to the ER process.

As pointed out previously (FYON *et al.* 2015), the ER process is notably influenced by three important genomic characteristics: (1) the occurrence and tightness of possible negative feedback regulatory loops, (2) the recombination rate between the regulated gene and the *cis*-regulator and (3) the intensity of purifying selection on the regulated gene.

It has been shown that stabilizing selection on expression levels significantly reduces the pace of the ER process (FYON *et al.* 2015), due to a cost of overexpression following enhancer strength escalation. Negative feedback regulatory loops minimize this effect, as they buffer the impact of enhancer strength escalation by maintaining more stable expression levels. It is thus easier for enhancer strength to escalate when there is a negative feedback regulatory loop. Hence, ER process is expected to proceed faster when gene expression is at least partially controlled through a negative feedback regulatory loop. In such cases, regulatory divergence between closely related species is expected to occur at a faster-than-neutral rate, as it is easier for proximal *cis*-regulatory sequences to increase in strength, and easier

for other regulatory sequences to compensate (transitory sub-optimal expression levels are less selected against). As genome-wide information on negative feedback regulatory circuit is currently not available, we used transcriptional noise as a proxy for the presence of negative feedback loops (BECSKEI and SERRANO 2000; THATTAI and VAN OUDENAARDEN 2001). We used a normalized measure of transcriptional noise, *F**, estimated from single-cell transcriptomes of the domestic mouse (Barroso and Dutheil, in prep). When *F** is low, gene expression is very stable between cells, and is probably regulated through tight negative feedback regulatory loops. On the contrary, when *F** is high, gene expression is noisy, and is probably not regulated through negative feedback regulatory loops. Consequently, we can make a first prediction: the ER process is expected to create a negative correlation between regulatory divergence and *F**. Indeed, when *F** is low, tight negative feedback regulatory loops are expected to provide room for *cis*-regulatory strength escalation and subsequent coevolution, favoring fast regulatory divergence among species. On the contrary, when *F** is high, there is less room for *cis*-regulatory strength escalation, and thus less regulatory divergence. Importantly, a scenario where sequence divergence is only governed by purifying selection, would lead to the opposite pattern: genes under strong purifying selection typically display lower promoter divergence and a more stable expression (low *F**). This is in particular reflected in the positive correlation of *Ka / Ks* ratios and *F** in protein coding genes (Barroso and Dutheil, in prep.). A negative correlation of *F** and sequence divergence may however be expected under neutral evolution and stabilizing selection. In cases of tight negative feedback regulatory loops (low *F**), any *cis*-regulatory change should have little impact on expression levels, and thus undergo near-neutral evolution. When such loops are loose or absent (high *F**), those regulatory changes impact more expression levels, and should be selected against by stabilizing selection on expression levels. Divergence is then smaller than what is expected under neutral evolution. Overall, the ER process is not the only evolutionary force that may lead to a negative correlation between regulatory divergence and *F**. We thus cannot use this correlation alone as a specific footprint of the ER process.

The rate of *cis*-regulatory strength escalation depends strongly on the intensity of the purifying selection on the regulated gene. Indeed, if there's little purifying selection, there is little advantage for stronger enhancers to get associated with more viable gene copies,

simply because those gene copies are not so much more viable. The indirect selection favoring stronger enhancers directly stems from the purifying selection on the gene. More intense purifying selection is thus expected to promote faster ER process, and thus wider regulatory divergence. We predict that the ER process produces a positive correlation between gene purifying selection and regulatory divergence. This is highly unexpected under a more classic hypothesis that would involve background selection. In this case, regions associated with genes under more purifying selection should diverge less than neutral, as their coalescence time before the speciation split is expected to be reduced by background selection. This effect is however only expected to be significant for recently diverged species. In long-diverged species, mutations occurring after the split outnumber by far those that occurred between split time and neutral coalescence time, making this effect negligible. However, correlations between selection forces may produce a signal of higher enhancer divergence at higher gene purifying selection. Indeed, chances are that stabilizing selection on expression levels and purifying selection on gene sequences are positively correlated: an important functional gene sequence might have been selected to be more stably expressed. In those cases, chances are that gene expression levels are tightly controlled through negative feedback loops to avoid any detrimental misexpression. Genes with strong purifying selection may be preferentially regulated through negative feedback loops. The presence of these regulatory feedback loops, in turn, may have important effects on *cis*-regulatory changes. First, the loop itself must be in part controlled by *cis*-regulatory elements. These sequences are likely to be strongly constrained, as mutations disrupting this loop are likely to be deleterious. Hence, with this type of mutation in regulatory regions, we expect lower regulatory divergence for genes with stronger purifying selection. However, the opposite prediction can be made considering *cis*-regulatory mutations that do not alter the regulatory loop, but e.g. change interaction between the transcription factors and enhancers. Because of the loop, these mutations will have little impact on phenotype (mostly regulated through the feedback loop) and will spread mostly neutrally. Hence, paradoxically, genes under strong purifying selection would be expected to show near neutral and thus stronger divergence at their *cis*-regulatory sequences compared to genes with no feedback loops and weak purifying selection. Overall, it is difficult to determine *a priori* whether gene with stronger purifying selection will exhibit stronger of weaker divergence. In any case, observing a positive correlation between regulatory divergence and

the strength of purifying selection on the corresponding gene may not unequivocally point towards the ER process. It could simply reflect that genes with stronger purifying selection are more likely to have loops, and that most mutations in their regulatory regions do not impact the loop itself.

The ER process results from the beneficial association between stronger enhancers and more viable gene. Recombination between *cis*-regulatory sequences and regulated genes is thus a crucial parameter: the rate of the ER process decreases for *cis*-regulators that are genetically more distant to the regulated gene (FYON *et al.* 2015). At high genetic distance, the ER process cannot even take place. The ER process is therefore expected to create a negative correlation between regulatory divergence and genetic distance between the gene and the regulatory sequence. Such a correlation is much more difficult to explain with alternative scenarios, in absence of the ER process. For instance, background selection is expected to be stronger for more tightly linked regulatory sequences, which should lead to the inverse pattern: increased divergence at larger recombination distance. As we explained above, background selection reduces coalescence time, and thus reduces neutral divergence. Hence, only *ad-hoc* hypothesis could explain closer regulatory sequences diverging more. For example, one could imagine that distant regulators are functionally more important (even though what is so far known about gene regulation mechanisms leads us to think otherwise), and thus more selectively constrained than regulators close to genes. However, even with this kind of *ad-hoc* hypotheses, regulatory sequences diverge at best neutrally. Indeed, background selection as well as stabilizing selection tends to reduce regulatory divergence. Only ER process, and rare events of directional selection, can create patterns of regulatory divergence faster than neutral. Directional selection, however, is not expected to create genome-wide average patterns of regulatory divergence stronger at close enhancers than at distant enhancers.

Overall, the ER process is expected to leave specific footprints in the genome. It should lead to: (1) stronger regulatory divergence for proximal enhancers than for distant enhancers; and (2) stronger-than-neutral divergence of proximal enhancers. This pattern is expected to be more pronounced in genes exhibiting low F^*, high purifying selection and low local recombination rates.

Methods

Measure of transcriptional noise

Single-cell transcriptome from Sasagawa et al. (2013) was used in order to compute variance in expression for each mouse cell. In order to compute transcription noise, the variance was plotted against the mean gene expression computed over all cells. The *F** measure was defined for each gene as the observed variance divided by the predicted value of the logarithmic regression. *F** is therefore a normalized measure of noise, independent of the mean expression. It has been calculated for 13,600 genes with sufficient data (Barroso and Dutheil, in prep).

Test of increased Mouse-Rat divergence in proximal regulatory regions

Sequence and annotations from Ensembl release 83 were used. Gene Feature Format (GFF) annotation files from the Ensembl website (accessed 17/12/15) were processed using dedicated perl scripts using the BioPerl GFF parser and tools. The genome tools package (GREMME *et al.* 2013) was used to generate intron coordinates. The Compara mouse-rat pairwise genome alignment from Ensembl was processed using MafFilter in order to extract alignment regions from GFF files and compute sequence divergence in each region (DUTHEIL *et al.* 2014). The number of mutations (approximated by the number of mismatches in the corresponding part of the alignment) in regions of 300 bp upstream each gene were compared to (1) the number of mutations in regions 600 to 900 bp upstream each gene, and (2) the total number of mutations in all introns of each gene (intron divergence is used as a neutral divergence proxy). In order to test whether there is significantly more mutations in regions of 300 bp upstream genes, a binomial test was performed. We refer to this region as region 1, and to the compared region as region 2 (600-900 bp regulatory regions or introns). We note m_1 and m_2 the numbers of mutations and n_1 and n_2 the total number of aligned positions in regions 1 and 2, respectively. Under the null hypothesis of uniform distribution of mutations, the number of mutations in region 1 follows a binomial distribution of parameter size equal to n_1 and probability $p = (m_1 + m_2) / (n_1 + n_2)$. The probability of observing at least b mutations in region 1 is given by the formula:

$$\sum_{i=b}^{m1} \binom{m1}{i} p^i (1-p)^{m1-i}$$

All tests were corrected for multiple testing using the Benjamini-Hochberg method (BENJAMINI and HOCHBERG 1995).

Recombination rate

Local mean recombination rates were obtained for each gene by averaging recombination estimates in a 50 kb window centered on the gene. Recombination rates from (BRUNSCHWIG *et al.* 2012) were used after being converted from mm9 to mm10 reference genome using the UCSC LiftOver utility.

Ka / Ks

Ka / Ks accounts for gene purifying selection intensity. Values between mouse and rat were retrieved from the Ensembl 83 database using the Biomart interface. Only data from one-to-one orthologs were used, with confidence equal to 1.

Linear models

We used logistic regression models in order to assess the effect of transcriptional noise, *Ka / Ks* and recombination rate on the relative divergences of each gene. We set a false-discovery rate threshold of 10% and used the significance of the mutation repartition test as a response variable. A generalized linear model with binomial family and logit link was fitted, and a stepwise model selection procedure was conducted, initially including all three predictor variables *(F**, *Ka / Ks* and local recombination rates) as well as their pairwise and triple interactions.

Model M1 uses the proximal vs. distant divergence test. Stepwise model selection retains transcriptional noise as well as *Ka / Ks* as explanatory variables, without interaction. Residues independence was confirmed using the Ljung-Box test (p-value = 0.8838, no significant departure from independence).

Model M2 uses the proximal vs. introns divergence test. Stepwise model selection retains transcriptional noise as well as *Ka / Ks* and their pairwise interaction as explanatory

variables. Residues independence was confirmed using the Ljung-Box test (p-value = 0.2977, no significant departure from independence).

Results

Proximal vs. distant divergence test

First, we want to know if proximal (0 – 300 bp) *cis*-regulatory sequences have diverged significantly more than distant (600 – 900 bp) sequences. To do that, we perform a mutation repartition test with a false-discovery rate of 10%, which returns 1 when proximal sequences have diverged significantly more, 0 otherwise. From this, we fit M1, a generalized linear model with binomial family and logit link. In a nutshell, we look at the probability that a gene has more divergent proximal regulatory sequences (returns 1 after the mutation repartition test), depending on predictor variables selected through a stepwise model selection. Selected predictor variables are expression noise *F** and purifying selection intensity *Ka / Ks*. Results are:

	Estimate	**p-value**	
Intercept	-3.5972	< 2e-16	***
F*	-0.5838	0.00347	**
Ka / Ks	1.5651	0.06399	

Intercept is negative, which means that genes who have more divergent proximal regulatory sequences are rare. Indeed, we found only 149 such genes. Most of the time, distant sequences diverge more than proximal ones, which is in accordance with classic background selection theory. However, we also see that the effect of *F** is significant and negative, such that there are more chances to find that a gene has more divergent proximal sequences if *F** is low, that is if this gene has a rather stable expression profile. This is more in accordance with ER theory: cases of increased regulatory divergence in proximal sequences should occur preferentially at low *F**, when the ER process is expected to be stronger. Effect of *Ka / Ks* is not significant.

Proximal vs. introns divergence test

We now repeat the same statistical procedure, except that the mutation repartition test occurs between proximal regulatory sequences and introns. We assume that intron sequences evolve mostly neutrally (though this is not completely true, see Gazave et al. 2001), and use intron divergence as a proxy of neutral divergence. Selected predictor variables are *F**, *Ka / Ks* and their interaction. Results are:

	Estimate	**p-value**	
Intercept	-3.2809	<2e-16	***
F*	-0.3720	0.0374	*
Ka / Ks	-2.0737	0.1174	
F* x Ka / Ks	1.4279	0.0334	*

Again, the intercept is negative, genes that have proximal regulatory sequences that diverge more than intron sequences are rare: we found only 226 of such genes. This means that proximal regulatory sequences usually diverge less than introns, meaning less than neutral. This is probably due to widespread stabilizing selection. However, the effect of *F** is again significant and negative: a gene will have more chance to have proximal regulatory sequences that diverge more than introns if it has a low *F**, that is a rather stable expression profile. Stronger-than-neutral proximal regulatory divergence tends to occur at low *F**, which is in accordance with ER process, though the effect is small and merely significant. Effect of *Ka / Ks* is not significant. Interaction effect of *F** and *Ka / Ks,* on the other hand, is significant and positive. This means the relationship between more-than-neutral divergence and low *F** works better at low *Ka / Ks*, where the ER process is expected to be stronger.

Genes with ER process genomic footprint

Most of genes do not have significantly higher divergence in proximal regulatory regions than in distant ones or introns. This is probably due to widespread background and stabilizing selection, which limits regulatory divergence especially for proximal sequences. However, applying a 10% false-discovery rate for the mutation repartition tests, we managed to find 149 genes with regulatory divergence higher for proximal than distant regulatory sequences, and 226 genes with proximal regulatory divergence higher than intron divergence. There is little overlap between the two groups, as we only found 27 genes that verify the two conditions, but this overlap is more than expected at random (Fisher's exact test: *p-value* < 2.2e-16). Despite widespread background and stabilizing selection in regulatory evolution, we found a few genes that exhibit ER process specific genomic footprint: they return 1 on both mutation repartition tests.

Discussion

Using the results from our models to establish predictions, we manage to define a genomic footprint of regulatory divergence that can only be explained by the ER process. This footprint is characterized by proximal *cis*-regulatory sequence divergence being stronger than divergence of distant *cis*-regulatory sequences, and stronger than neutral divergence (we use intron divergence as a proxy of neutral divergence).

Using divergence data between *Mus musculus* and *Rattus norvegicus*, we found at least 27 genes with signature of the ER process.

Weakness of the signal may have many explanations. The ER process proceeds from indirect selection on regulatory sequences that is rather weak: of the order of magnitude of the mutation rate on the gene (FYON *et al.* 2015). We thus do not expect a strong signal. Moreover, this small positive selection tends to be obscured by widespread negative selection signals, like background selection or stabilizing selection signals. Stabilizing selection seems to be particularly widespread (LUDWIG *et al.* 2000; DENVER *et al.* 2005), and tends to slow down sequence divergence (FAY and WITTKOPP 2008) and – as our models also showed – ER process' rate (FYON *et al.* 2015).

For example, there are 122 genes displaying proximal enhancer divergence stronger than distant enhancer divergence but smaller than neutral divergence. It is possible that the regulatory sequences of these genes undergo ER process such that proximal enhancer divergence is stronger than distant enhancer divergence. However, some stabilizing selection may also act such that proximal enhancer divergence is smaller than neutral divergence.

Weakness of the signal may also be due to the limitations of our method. First, we use intron divergence as a proxy of neutral divergence. However, we know that introns do not evolve completely neutrally as, among other functions, they impact alternative splicing and play some role in regulatory networks (GAZAVE *et al.* 2001; CHOREV and CARMEL 2012). If intron evolution is primarily affected by stabilizing selection, introns diverge less than neutral, such that proximal enhancers that diverge more than introns may actually diverge less than neutral: some of the 27 genes may actually not display the signal we are interested in. On

the other hand, if introns are primarily affected by positive selection, then there are more than 226 genes whose regulatory sequences diverge more than neutrally, such that in the end our signal may concern more than 27 genes. Usually, pseudogenes are used as a proxy of neutral divergence (FAY and WITTKOPP 2008), but we could not identify pseudogenes from Ensembl data set. An alternative to pseudogenes is using synonymous sites (HALLIGAN *et al.* 2004): synonymous mutations lead to the same amino-acid sequence, and thus presumably do not lead to any phenotypic change. However, evidence accumulates that suggests synonymous sites actually may not evolve neutrally, mainly due to impacts of synonymous mutations on mRNA splicing and stability (HALLIGAN *et al.* 2004; AMORÓS-MOYA *et al.* 2010). Eventually, we chose to use introns as the neutral proxy but, in a more thorough study of ER process signal, one will have to use several proxies for robust neutral divergence estimation.

The signal we are looking for is also blurred by the multiple correlations involving selection and negative feedback loops. Indeed, *Ka / Ks* is positively correlated with *F**. This is because intensity of purifying selection is positively correlated with intensity of stabilizing selection, which is also positively correlated with the presence of negative feedback loops. Regulatory sequences associated with genes under high purifying selection are expected to diverge quickly due to the ER process, but also undergo a strong stabilizing selection that slows down divergence. However, such sequences might also be associated with tighter negative feedback loops, which may allow for accelerated divergence. As we can see, there are many selective effects interfering and impacting differently regulatory divergence, many of which are correlated with each other, such that it may be difficult to decipher the main drivers of regulatory sequence evolution. The work done here would benefit from being replicated on multiple species. It would also benefit from more precise knowledge of gene regulatory networks. In particular, there is a strong need for better annotation of negative feedback loops. There is also a strong need for better annotation of *cis*-regulatory sequences. Indeed, because it is difficult to point regulatory sequences, we did not work on *cis*-regulatory sequences *per se*, but on non-coding regions upstream of genes that are known to be enriched in regulatory sequences.

Despite all these limitations, we managed to find a genomic footprint specific to ER process. We provide the first evidence for a process we previously theorized (FYON *et al.* 2015). This adds up with previous clues found in the literature. For example, we predicted that, in F1

hybrids of closely-related species differing by their reproduction mode (selfing vs. outcrossing), genes from the outcrossing parent should be more expressed (FYON *et al.* 2015). This pattern was found between *A. thaliana* and *A. lyrata* by He et al. (2012). However, they could not make reciprocal crosses, such that their results may be explained by parent-of-origin effects. Metzger et al. (2016) found that random *cis*-regulatory mutations preferentially decrease expression levels, whereas random *trans*-regulatory mutations preferentially increase expression levels, a pattern we also predicted (FYON *et al.* 2015). Overall, ER process seems to have the potential to explain several empirical patterns. Because it is very general, and only requires diploidy to operate, this process may have occurred since the origin of eukaryotes more than a billion years ago. As we saw in the previous Chapters, it may have various and far-reaching consequences on regulatory architecture, sequence divergence, homologous chromosome relative expression and sex chromosome evolution. It may thus have profoundly influenced diploids' regulatory sequences. A whole new segment of regulatory evolution is being uncovered, which offers exciting perspectives. It may be therefore important to further test predictions associated to the ER process. In particular, the molecular approach developed here could be improved and applied to many other species pairs, which might overcome some of the statistical limitations that are intrinsic to the detection of weak selection.

References

ABZHANOV A., PROTAS M., GRANT B. R., GRANT P. R., TABIN C. J., 2004 Bmp4 and morphological variation of beaks in Darwin's finches. Science **305**: 1462–5.

AMORÓS-MOYA D., BEDHOMME S., HERMANN M., BRAVO I. G., 2010 Evolution in Regulatory Regions Rapidly Compensates the Cost of Nonoptimal Codon Usage. Mol. Biol. Evol. **27** : 2141–2151.

BECSKEI A., SERRANO L., 2000 Engineering stability in gene networks by autoregulation. Nature **405**: 590–593.

BENJAMINI Y., HOCHBERG Y., 1995 Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. J. R. Stat. Soc. Ser. B **57**: 289–300.

BRUNSCHWIG H., LEVI L., BEN-DAVID E., WILLIAMS R. W., YAKIR B., SHIFMAN S., 2012 Fine-Scale Maps of Recombination Rates and Hotspots in the Mouse Genome. Genetics **191**: 757 LP – 764.

CHOREV M., CARMEL L., 2012 The Function of Introns. Front. Genet. **3**: 55.

COOPER T. F., 2003 Parallel changes in gene expression after 20,000 generations of evolution in Escherichia coli. Proc. Natl. Acad. Sci. U. S. A. **100**: 1072–1077.

DENVER D. R., MORRIS K., STREELMAN J. T., KIM S. K., LYNCH M., THOMAS W. K., 2005 The transcriptional consequences of mutation and natural selection in Caenorhabditis elegans. Nat. Genet. **37**: 544–8.

DUTHEIL J. Y., GAILLARD S., STUKENBROCK E. H., 2014 MafFilter: a highly flexible and extensible multiple genome alignment files processor. BMC Genomics **15**: 53.

FAY J. C., MCCULLOUGH H. L., SNIEGOWSKI P. D., EISEN M. B., 2004 Population genetic variation in gene expression is associated with phenotypic variation in Saccharomyces cerevisiae. Genome Biol. **5**: 1–14.

FAY J. C., WITTKOPP P. J., 2008 Evaluating the role of natural selection in the evolution of gene regulation. Heredity. **100**: 191–9.

FEREA T. L., BOTSTEIN D., BROWN P. O., ROSENZWEIG R. F., 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc. Natl. Acad. Sci. U. S. A. **96**: 9721–9726.

FYON F., CAILLEAU A., LENORMAND T., 2015 Enhancer Runaway and the Evolution of Diploid Gene Expression. PLoS Genet **11**: e1005665.

GAZAVE E., FERNANDO O., NAVARRO A., 2001 The Evolution of Introns in Human Genes. In: *eLS*, John Wiley & Sons, Ltd.

GILAD Y., OSHLACK A., SMYTH G. K., SPEED T. P., WHITE K. P., 2006 Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature **440**: 242–245.

GREMME G., STEINBISS S., KURTZ S., 2013 GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. IEEE/ACM Trans. Comput. Biol. Bioinforma. **10**: 645–56.

HALLIGAN D. L., EYRE-WALKER A., ANDOLFATTO P., KEIGHTLEY P. D., 2004 Patterns of Evolutionary Constraints in Intronic and Intergenic DNA of Drosophila. Genome Res. **14**: 273–279.

HE F., ZHANG X., HU J., TURCK F., DONG X., GOEBEL U., BOREVITZ J., MEAUX J. DE, 2012 Genome-wide Analysis of Cis-regulatory Divergence between Species in the Arabidopsis Genus. Mol. Biol. Evol. **29**: 3385–3395.

HSIEH W.-P., CHU T.-M., WOLFINGER R. D., GIBSON G., 2003 Mixed-Model Reanalysis of Primate Data Suggests Tissue and Species Biases in Oligonucleotide-Based Gene Expression Profiles. Genetics **165**: 747 LP – 757.

KHAITOVICH P., HELLMANN I., ENARD W., NOWICK K., LEINWEBER M., FRANZ H., WEISS G., LACHMANN M., PÄÄBO S., 2005 Parallel Patterns of Evolution in the Genomes and Transcriptomes of Humans and Chimpanzees. Science (80-.). **309**: 1850 LP – 1854.

KHAITOVICH P., WEISS G., LACHMANN M., HELLMANN I., ENARD W., MUETZEL B., WIRKNER U., ANSORGE W., PÄÄBO S., 2004 A Neutral Model of Transcriptome Evolution. PLoS Biol. **2**: e132.

LO H. S., WANG Z., HU Y., YANG H. H., GERE S., BUETOW K. H., LEE M. P., 2003 Allelic variation in gene expression is common in the human genome. Genome Res. **13**: 1855–62.

LUDWIG M. Z., BERGMAN C., PATEL N. H., KREITMAN M., 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**: 564–7.

METZGER B. P. H., DUVEAU F., YUAN D. C., TRYBAN S., YANG B., WITTKOPP P. J., 2016 Contrasting frequencies and effects of cis- and trans-regulatory mutations affecting gene expression. Mol. Biol. Evol. .

NUZHDIN S. V, WAYNE M. L., HARMON K. L., MCINTYRE L. M., 2004 Common Pattern of Evolution of Gene Expression Level and Protein Sequence in Drosophila. Mol. Biol. Evol. **21** : 1308–1317.

OLEKSIAK M. F., CHURCHILL G. A., CRAWFORD D. L., 2002 Variation in gene expression within and among natural populations. Nat. Genet. **32**: 261–6.

RAYMOND M., CHEVILLON C., GUILLEMAUD T., LENORMAND T., PASTEUR N., 1998 An overview of the evolution of overproduced esterases in the mosquito Culex pipiens. Philos. Trans. R. Soc. B Biol. Sci. **353**: 1707–1711.

RIFKIN S. A., HOULE D., KIM J., WHITE K. P., 2005 A mutation accumulation assay reveals a broad capacity for rapid evolution of gene expression. Nature **438**: 220–3.

RIFKIN S. A., KIM J., WHITE K. P., 2003 Evolution of gene expression in the Drosophila melanogaster subgroup. Nat Genet **33**.

SASAGAWA Y., NIKAIDO I., HAYASHI T., DANNO H., UNO K. D., IMAI T., UEDA H. R., 2013 Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. **14**: 1–17.

SHAPIRO M. D., MARKS M. E., PEICHEL C. L., BLACKMAN B. K., NERENG K. S., JÓNSSON B., SCHLUTER D., KINGSLEY D. M., 2004 Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. Nature **428**: 717–23.

THATTAI M., OUDENAARDEN A. VAN, 2001 Intrinsic noise in gene regulatory networks. Proc. Natl. Acad. Sci. **98** : 8614–8619.

UDDIN M., WILDMAN D. E., LIU G., XU W., JOHNSON R. M., HOF P. R., KAPATOS G., GROSSMAN L. I., GOODMAN M., 2004 Sister grouping of chimpanzees and humans as revealed by

genome-wide phylogenetic analysis of brain gene expression profiles. Proc. Natl. Acad. Sci. United States Am. **101** : 2957–2962.

WEIRAUCH M. T., HUGHES T. R., 2010 Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. **26**: 66–74.

WHITEHEAD A., CRAWFORD D. L., 2006a Variation within and among species in gene expression: raw material for evolution. Mol. Ecol. **15**: 1197–211.

WHITEHEAD A., CRAWFORD D. L., 2006b Neutral and adaptive variation in gene expression. Proc. Natl. Acad. Sci. **103**: 5425–5430.

YANAI I., GRAUR D., OPHIR R., 2004 Incongruent Expression Profiles between Human and Mouse Orthologous Genes Suggest Widespread Neutral Evolution of Transcription Control. Omi. A J. Integr. Biol. **8**: 15–24.

Conclusion Générale

Résumé des épisodes précédents

Dans ce mémoire, nous avons pu montrer à l'aide d'outils théoriques qu'une sélection indirecte jusqu'alors jamais évoquée dans la littérature pourrait agir au niveau de certaines séquences *cis*-régulatrices et se propager à travers les réseaux de régulation via des coévolutions. Cette sélection indirecte découle des déséquilibres d'expression entre copies homologues résultant du polymorphisme des séquences *cis*-régulatrices. Dans le cas général (Chapitres I et II), cette sélection indirecte privilégie les promoteurs plus forts (augmentant davantage les niveaux d'expression) que leurs homologues. Cette sélection indirecte devrait alors aboutir à une augmentation continue de la force des promoteurs concernés, autant qu'il est possible. Nous avons appelé ce processus *ER process*.

Dans le cas de lignées clonales (ou presque clonales), cependant, un tout autre patron apparaît (Chapitre III) : on s'attend cette fois à ce que l'expression des gènes deviennent progressivement haploïde. Le processus de fond est le même : les promoteurs forts s'associent aux allèles les plus viables du gène, tandis que les promoteurs faibles s'associent aux allèles les moins viables du gène. Mais l'absence de brassage génétique empêche un promoteur fort de se fixer entièrement : sans conversion génique, ni sexe ni recombinaison, le promoteur fort ne pourra jamais s'implanter sur le chromosome homologue de celui où il est apparu. Les individus étant « condamnés » à l'hétérozygotie, une ségrégation se fait, les promoteurs forts s'accumulant à proximité d'une copie du gène purgée de mutations délétères tandis que les promoteurs faibles s'accumulent à proximité de la copie homologue. Cette copie homologue accumule pseudo-neutralement des allèles délétères car elle ne participe plus ou presque plus à l'expression totale du gène. Nous appelons ce processus *ED process.*

Nous avons étudié un autre cas particulier : le cas des chromosomes sexuels (Chapitre IV). On a considéré une paire de jeunes chromosomes sexuels ayant cessé de recombiner. Ces chromosomes étant isolés génétiquement, ils subissent l'*ED process*. De façon importante, ici, c'est toujours le chromosome Y qui accumule les promoteurs faibles et les allèles délétères, tandis que le chromosome X accumule toujours les prmoteurs forts et les allèles viables. Ceci est dû à l'asymétrie des chromosomes X et Y. Le chromosome X étant à l'état homozygote chez les femelles, il ne peut accumuler des promoteurs faibles et des allèles

délétères. L'*ED process* entre chromosomes sexuels reproduit ainsi l'évolution des chromosomes sexuels, avec dégénérescence et extinction du chromosome Y et sur-expression du chromosome X. On a vu également que les différents modes de compensation de dosage, pour rétablir des niveaux d'expression des chromosomes sexuels similaires à ceux des autosomes, sont compatibles avec l'*ED process*.

Limites des approches utilisées

Comme dans tout travail de modélisation, nous nous sommes efforcés de simplifier des mécanismes naturels, et ce en vue de voir apparaître les grandes forces structurant notre système d'étude. Cette simplification fait appel à un certain nombre d'hypothèses dont la remise en cause est susceptible d'altérer significativement les résultats. Il nous faut donc prendre le temps de discuter ces hypothèses, afin de bien en comprendre les implications.

La simplification la plus évidente que nous avons considérée est celle du réseau de régulation. Nous avons la plupart du temps considéré un seul élément régulateur isolé, éventuellement deux, mais nous n'avons jamais considéré un réseau régulateur dans toute sa complexité. Nous avons également considéré que les mutations affectent uniquement les niveaux d'expression. Ceci dans un souci de clarté, afin de se concentrer sur les conséquences sélectives de la variation d'expression relative de copies homologues. Dans la réalité, il n'est pas assuré qu'une mutation affectant la force d'un promoteur, par exemple, ne puisse avoir d'effets pléiotropes par ailleurs. Les promoteurs attirant plusieurs facteurs de transcription, chacun en relation avec différents facteurs et co-facteurs, des effets pléiotropes sont à prévoir, sur les niveaux d'expression d'autres gènes, mais aussi sur d'autres caractéristiques d'expression que les seuls niveaux d'expression. Cependant, la structure modulaire des séquences *cis*-régulatrices est considérée par beaucoup comme réduisant la pléiotropie des mutations sur ces séquences, chaque module contrôlant un aspect particulier des profils d'expression (WRAY 2007; WITTKOPP and KALAY 2012). Il n'est donc pas exclu de penser que des mutations soient susceptibles de modifier uniquement des niveaux d'expression, sans effet pléiotrope majeur. De plus, l'existence de réseaux équivalents, différents mais menant au même profil d'expression, peut permettre aux séquences régulatrices de co-évoluer de sorte à compenser d'éventuels effets pléiotropes.

Nous avons régulièrement incorporé la recombinaison dans notre modèle, mais jamais la conversion génique, un autre aspect important de la méiose (voir Encart 2 dans l'Introduction Générale). La conversion génique a un double effet. Elle casse les associations génétiques en transformant une copie en sa copie homologue. Elle agit aussi sur l'hétérozygotie : en transformant des hétérozygotes en homozygotes, elle tend à diminuer l'hétérozygotie d'une population. Il est donc à prévoir que des taux suffisamment élevés de

conversion génique puissent diminuer significativement la sélection indirecte sur les promoteurs. Pour l'*ER process*, on s'attend ainsi à ce que la conversion génique ralentisse l'escalade des forces des promoteurs, jusqu'à éventuellement la stopper si l'hétérozygotie et les associations génétiques favorables sont trop faibles. Pour l'*ED process*, on s'attend à ce que la conversion génique agisse comme la recombinaison avec les centromères dans les cas d'automixie en fusion centrale : elle rompt l'isolement génétique des chromosomes homologues et crée des homozygotes, et tend donc à limiter la divergence des séquences homologues.

Le contexte sélectif des séquences régulatrices et des gènes a également été simplifié : nous n'avons considéré qu'une sélection purifiante sur le gène et éventuellement une sélection stabilisante sur les niveaux d'expression du gène. Nous avons pu vérifier que l'*ER process* fonctionne aussi bien sous un régime de mutations bénéfiques dominantes que sous un régime de mutations délétères récessives. Il reste cependant de nombreux contextes sélectifs inexplorés, tant au niveau du gène (superdominance, sous-dominance, sélection stabilisante) qu'au niveau des séquences régulatrices (sélection directionnelle, sélection stabilisante fluctuante). S'il est probable que, dans la plupart des cas, une sélection indirecte s'appliquera toujours aux séquences régulatrices, il serait intéressant de comparer les vitesses d'évolution des séquences régulatrices dues à cette sélection indirecte avec les vitesses d'évolution dues à d'autres sélections s'appliquant aux séquences régulatrices.

Limites des processus

Il convient de bien comprendre que, dans cette thèse, le choix des valeurs des paramètres a été fait de sorte à illustrer le processus sélectif étudié et ses conséquences. Certaines gammes de paramètres sont explorées, de sorte que l'on puisse étudier les différents cas de figures. D'autres paramètres sont fixés, à des valeurs qui permettent soit de voir clairement le processus (forte sélection purifiante, taux de mutations importants), soit de limiter le temps de simulation (le processus est plus visible à de plus fortes tailles de population, mais les simulations sont trop longues).

L'analyse a montré que l'intensité de la sélection indirecte sur les séquences régulatrices est de l'ordre du taux de mutation sur le gène. Ce taux dépend des espèces et des gènes considérés, mais peut être faible, voire très faible. Cette faiblesse de l'intensité de la sélection peut être cependant compensée par des grandes tailles de population, de sorte que $Ns >> 1$.

On a aussi vu que la sélection stabilisante sur les niveaux d'expression tend à ralentir l'évolution des promoteurs, dans la mesure où cette évolution doit se faire en concordance avec l'évolution d'autres séquences régulatrices afin de maintenir des niveaux d'expression optimaux. Si, en l'absence de boucles de rétroaction négative, la sélection stabilisante est trop forte, l'*ER process* peut être trop lent pour être biologiquement significatif. A moins que les possibilités de coévolution du réseau soient suffisamment fines pour entraîner des niveaux d'expression transitoires suffisamment proches du niveau optimal. La sélection stabilisante semble être répandue au sein des réseaux de régulation, de même que la coévolution entre séquences régulatrices. La fréquence des boucles de rétroaction négative, en revanche, n'est pas connue.

Le taux de recombinaison entre la séquence *cis*-régulatrice et le gène régulé est un paramètre important du modèle, puisque la recombinaison rompt les associations génétiques à la base de la sélection indirecte sur les séquences *cis*-régulatrices. Comme nous l'avons montré pour l'*ER process*, il existe une fenêtre de recombinaison proche du gène au-delà de laquelle, quand le promoteur est situé trop loin, il n'est pas soumis à l'*ER process*. La limite de cette fenêtre dépend de l'intensité de la sélection indirecte sur les promoteurs, qui dépend elle-même des autres paramètres (taux de mutation, intensité de la sélection

purifiante sur le gène, taille de la population, etc). Il est donc difficile de dire quelles séquences *cis*-régulatrices d'organismes réels devraient être soumises au *ER process*. Cependant, il est possible que, chez certains gènes, la limite de la fenêtre est si faible que toutes les séquences *cis*-régulatrices du gène ou presque soient au-delà de la limite. En ce qui concerne l'*ED process*, nous avons vu qu'il est limité à une très forte isolation génétique des chromosomes homologues. On a ainsi vu que de très faibles taux de reproduction sexuée, ou de recombinaison avec les centromères en cas d'automixie avec fusion centrale pouvait empêcher l'*ED process* d'avoir lieu. La recombinaison mitotique, que nous n'avons pas prise en compte dans nos modèles, est susceptible de limiter de façon similaire l'*ED process* en cas d'apomixie.

L'association génétique entre promoteurs forts et gènes viables est un processus vraisemblablement faible la plupart du temps. Toutefois, cette limite est contrebalancée par son caractère général : ce processus doit être en cours depuis l'apparition des premiers diploïdes il y a plusieurs centaines de millions / milliards d'années. Après autant de temps, même un processus faible peut avoir le temps de mener à des modifications remarquables.

Principales conséquences génomiques

Gradient de force des promoteurs

Les changements induits par la sélection indirecte sur les *cis*-régulateurs sont nombreux. Le premier changement, immédiat, est bien sûr d'augmenter la force des promoteurs suffisamment proches des gènes. Cette augmentation dépend de la localisation des différentes séquences *cis*-régulatrices : plus un promoteur est près du gène, plus on s'attend à ce qu'il soit fort. Une seconde force sélective s'applique aux *cis*-régulateurs : la force de la sélection stabilisante, qui tend à faire coévoluer les différents régulateurs pour assurer des niveaux d'expression optimaux malgré l'augmentation de force des promoteurs proches des gènes. Comme nous l'avons montré dans le Chapitre II, cette coévolution peut se faire entre *cis*-régulateurs. Dans ce cas, plus un promoteur est loin du gène, moins il subira l'influence du *ER process* et plus il aura tendance à avoir un rôle compensateur pour l'augmentation des forces des promoteurs plus proches du gène. Comme illustré en Annexe, on s'attend donc à observer un gradient dans l'organisation des séquences *cis*-régulatrices, avec des promoteurs très activateurs proches des gènes, et des promoteurs plutôt inhibiteurs loin des gènes.

Une architecture régulatrice complexe

Comme nous l'avons fait remarqué dans le Chapitre I, l'*ER process* est susceptible de complexifier les réseaux de régulation. En effet, plusieurs voies sont possibles pour augmenter la force d'un promoteur. Certaines n'affectent pas la complexité du réseau : un promoteur peut ainsi subir une mutation qui augmente son attractivité pour un facteur de transcription activateur. D'autres, en revanche, augmentent la complexité du réseau. Ainsi, une séquence reconnue par un facteur de transcription activateur peut être dupliquée, de sorte que les régions *cis*-régulatrices attirent davantage de facteurs de transcription. Les séquences de reconnaissance des facteurs de transcription peuvent ainsi se multiplier, et éventuellement se différencier si l'attraction d'un nouveau facteur de transcription permet d'augmenter la force des promoteurs.

De façon intéressante, il y a toutes les chances pour que la complexité du réseau facilite l'*ER process*. En effet, face à la sélection stabilisante sur les niveaux d'expression, il est plus facile

de trouver des chemins de coévolution assurant un maintien des niveaux d'expression avec un grand nombre de régulateurs. On se retrouve donc dans un système qui s'auto-entretien : l'*ER process* favorise une complexité accrue du réseau, complexité qui à son tour favorise le processus de coévolution qui permet à l'*ER process* de continuer.

Les avis divergent concernant les raisons évolutives de la complexité des réseaux de régulations. Alors que certains affirment que cette complexité est absolument nécessaire aux mécanismes de régulation (LENSKI *et al.* 2003; CROMBACH and HOGEWEG 2008; JENKINS and STEKEL 2010), d'autres estiment qu'il est possible que cette complexité ne soit pas le résultat d'un processus adaptatif (SOYER and BONHOEFFER 2006; BIGGIN 2014). Etant donné les centaines de millions d'années dont a disposé l'*ER process*, et la boucle positive entre *ER process* et complexité du réseau, il est tout à fait envisageable de penser que l'*ER process* a pu participer à l'édification des réseaux de régulations complexes observés chez les eucaryotes (qui sont diploïdes pour une phase de leur cycle de vie à la différence des procaryotes chez qui ces régulations sont nettement plus simples).

Une divergence régulatrice accélérée

L'*ER process* pourrait bien compter pour une part significative des signaux de sélection positive observés dans les génomes, à côté des cas de sélection directionnelle. L'*ER process* agit comme une force qui favorise le remplacement des séquences régulatrices, qu'il s'agisse des promoteurs proches des gènes, ou des autres régulateurs qui coévoluent en compensation. L'*ER process* a donc toutes les chances de favoriser la divergence des séquences régulatrices entre espèces proches. Et ce d'autant plus que de nombreuses façons de compenser pour l'augmentation des forces des promoteurs proximaux peuvent exister. Si deux espèces proches subissent la même augmentation de la force des promoteurs proximaux, rien ne garantit que la coévolution avec les autres régulateurs prenne la même voie dans les deux cas. De sorte que, après spéciation, les réseaux de régulation de deux espèces proches peuvent être assez différents sans que l'on puisse voir de différences phénotypiques notables entre les deux espèces (WEIRAUCH and HUGHES 2010).

Il est reconnu que l'évolution des phénotypes et l'évolution des réseaux de régulation sont peu corrélées : des réseaux proches peuvent induire des phénotypes complètement différents, tandis que des réseaux différents peuvent être équivalents et mener au même

phénotype. Cette déconnection est généralement envisagé comme permettant une évolution neutre entre réseaux équivalents. Cette évolution neutre tourne à vide : les séquences changent, mais pas le phénotype. Pour cette raison, les séquences régulatrices peuvent diverger plus vite que les séquences codantes.

L'*ER process* amène à regarder ce processus d'un œil neuf. Il a en effet toutes les chances de jouer comme un moteur de cette évolution qui tourne à vide : la coévolution compensatrice remplace de nombreuses séquences pour s'assurer que les niveaux d'expression restent optimaux. Les organismes passent ainsi d'un réseau équivalent à l'autre, au fur et à mesure que la force des promoteurs proximaux augmente. L'*ER process* amène donc à prédire une évolution des séquences régulatrices plus rapide que celle induite par le remplacement neutre d'un réseau par un réseau équivalent, puisqu'il favorise sélectivement le remplacement des séquences régulatrices. L'*ER process* peut être vu comme une force sélective favorisant, parmi des réseaux équivalents, ceux avec des promoteurs proximaux les plus forts.

Haploidisation dans les lignées clonales

Lorsque la reproduction des organismes considérés est clonale ou presque clonale, nous n'assistons pas à une augmentation générale de la force des promoteurs proximaux, mais à une divergence des séquences régulatrices. Les séquences *cis*-régulatrices proches d'un gène et présentes sur un certain chromosome augmentent en force, tandis que les mêmes séquences sur le chromosome homologue diminuent en force. Toute l'expression du gène est progressivement monopolisée par une copie qui produit tous les ARN messagers correspondants, tandis que l'autre copie est éteinte.

Cette haploïdisation des gènes est attendue chez les organismes se reproduisant par mitose (apomixie). Chez de tels organismes, l'impact du processus dépendra des mêmes paramètres que dans les populations sexuées (intensité de la sélection purifiante, taux de mutation), mais aussi de la recombinaison mitotique, qui n'est pas prise en compte dans nos modèles. La recombinaison mitotique a vraisemblablement les mêmes effets sur l'intensité du processus que la recombinaison méiotique pour l'*ER process* : elle dissocie les associations génétiques favorables et ralentit, voire empêche, le processus. Nos modèles ne prennent pas en compte non plus la conversion génique, comme nous l'avons déjà mentionné. La

conversion génique est susceptible de réduire l'isolement des chromosomes homologues, et de permettre aux promoteurs plus forts d'envahir tous les chromosomes. Des taux faibles de conversion génique sont ainsi susceptibles de transformer l'haploïdisation attendue en *ER process*. Des taux très élevés de conversion génique devraient même être capable d'arrêter l'*ER process* en produisant trop d'homozygotes.

L'haploïdisation de l'expression est également attendue dans certaines populations se reproduisant par méiose. Ces populations doivent se reproduire presque exclusivement par automixie (les zygotes sont produits à partir d'une seule méiose). Deux types seulement de reproduction automictique entraîne une haploïdisation : la fusion centrale (fusion de deux gamètes provenant de deux secondes divisions méiotiques différentes) et le doublement pré-méiotique (doublement du matériel génétique avant la méiose). La recombinaison méiotique entre le gène et ses *cis*-régulateurs, comme la recombinaison mitotique pour les populations apomictiques, ralentit voire empêche le processus. Plus important cependant, en cas d'automixie avec fusion centrale, la recombinaison méiotique entre le gène et ses *cis*-régulateurs d'un côté et les centromères de l'autre a le même effet que la conversion génique. A des taux faibles, elle rompt l'isolement des chromosomes homologues, empêchant ainsi l'haploïdisation, et mène à une escalade des forces des promoteurs. A des taux élevés, elle produit trop d'homozygotes et empêche aussi l'escalade d'avoir lieu. On s'attend donc à ce que l'haploïdisation ait lieu pour des gènes plutôt proches des centromères.

Evolution des chromosomes sexuels

Nous avons également montré que l'*ED process* pouvait être à l'origine, ou avoir contribué, à l'évolution des chromosomes sexuels. En effet, l'haploïdisation de l'expression observée pour les lignées clonales est également valable pour les chromosomes sexuels chez le sexe hétérogame (XY ou XZ), du fait de la transmission particulière des chromosomes sexuels et du fait de l'arrêt de la recombinaison entre les chromosomes X et Y (ou W et Z). Cette haploïdisation est à l'origine à la fois de l'accumulation de mutations délétères sur les chromosomes Y (W), de l'extinction des chromosomes Y (W), et de l'augmentation de l'expression des chromosomes X (Z). Par la suite, si l'augmentation de l'expression des chromosomes X (Z) n'est pas limitée aux mâles (femelles), et s'il y a une pression de

sélection pour maintenir un dosage optimal entre expression des autosomes et expression des chromosomes sexuels, un mécanisme de compensation de dosage se met en place chez les femelles (mâles) pour diminuer l'expression global des chromosomes X (Z).

La sélection indirecte sur les séquences régulatrices à l'origine de l'*ED process* est faible (sauf en grande population), de l'ordre du taux de mutation sur le gène. Elle est cependant en action dès l'arrêt de la recombinaison entre chromosomes X et Y (Z et W). Cette sélection pourrait très bien expliquer les évolutions convergentes des chromosomes sexuels. De façon intéressante, la plupart des théories classiquement invoquées pour expliquer l'évolution des chromosomes sexuels (cliquet de Muller, interférences sélectives), semblent efficace surtout à faible taille de population. Notre théorie, en revanche, est plus efficace à grande taille de population, là où l'on s'attend à ce que la fixation d'allèles délétères par effets d'interférence sélective classiquement considérés soit relativement lente.

Des observations empiriques compatibles avec l'*ER process*

Les résultats obtenus par les modèles permettent de formuler un certain nombre de prédictions. Ainsi, par exemple, on prédit une architecture des réseaux plus compliquées que nécessaires pour assurer une bonne régulation de l'expression des gènes. Toutefois, ces prédictions ne sont pas nécessairement exclusives à notre processus. Par exemple, l'*ER process* n'est pas la seule force évolutive pouvant expliquer une architecture complexe. Cette complexité pourrait être due au hasard et avoir évolué neutralement, ou au contraire pourrait être adaptative.

Nous prédisions dans le Chapitre I que les mutations sur les *cis*- et les *trans*-régulateurs devraient être biaisées. En effet, après un long temps d'évolution, on s'attend à ce que certains *cis*-régulateurs aient atteint une force quasi-maximale, tandis que des *trans*-régulateurs auraient plutôt évolués pour diminuer les niveaux d'expression. Des mutations aléatoires sur de tels *cis*-régulateurs auraient donc plutôt tendance à diminuer les niveaux d'expression (peu de mutations sont capables d'augmenter encore plus les niveaux d'expression). A l'inverse, des mutations aléatoires sur de tels *trans*-régulateurs auraient plutôt tendance à augmenter les niveaux d'expression. C'est exactement ce qu'ont observé Metzger et ses collègues (METZGER *et al.* 2016).

Une autre prédiction concerne la vitesse de l'*ER process* en fonction des modes de reproduction. En effet, on s'attend à ce que l'*ER process* soit plus rapide lorsque la population se reproduit entièrement par allofécondation que lorsqu'elle se reproduit par autofécondation. On prédit donc que, globalement, les promoteurs proximaux d'espèces allogames soient plus forts que les promoteurs proximaux d'espèces autogames. Pour vérifier cette prédiction, il est possible de faire des croisements d'espèces proches allogames et autogames. Chez l'hybride, les facteurs de transcription des deux espèces parentes sont partagés. D'éventuelles différences d'expression chez certains gènes entre copies venant du parent allogame et copies venant du parent autogame ne peuvent alors être expliquées que par des différences dans les forces des promoteurs. Il faut donc montrer, pour confirmer notre théorie, que les copies venant du parent allogame sont en général plus exprimées que les copies venant du parent autogame. Ce résultat a été obtenu par He et ses collaborateurs (HE *et al.* 2012), qui ont croisé des plants d'*Arabidopsis thalliana* (autogame) avec des plants

d'*Arabidopsis lyrata* (allogame). Le même type de croisement a été effectué par Steige et ses collègues entre *Capsella rubella* (autogame) et *Capsella grandiflora* (allogame) (STEIGE *et al.* 2015). Cette dernière étude a obtenu des résultats opposés, montrant en général une plus forte expression des copies provenant du parent autogame. Cependant, ces deux études n'ont pas été capables de réaliser des croisements dans les deux sens (père autogame et mère allogame, et vice-versa). Il n'est donc pas possible de séparer les effets 'mode de reproduction du parent' des effets 'sexe du parent'. Une étude complète, avec des croisements dans les deux sens et des espèces proches mais aussi divergentes dans leurs modes de reproduction est nécessaire avant de pouvoir sérieusement tester de cette manière l'*ER process*. Il est à noter toutefois que l'influence des modes de reproduction, hors lignées clonales, sur l'*ER process* se fait exclusivement à travers l'hétérozygotie. Il peut donc s'avérer intéressant de se focaliser sur des espèces proches présentant des taux d'hétérozygotie particulièrement contrastés.

Une première confirmation empirique

L'une des prédictions que l'on a pu faire grâce à nos modèles concerne la vitesse de divergence des séquences régulatrices entre espèces proches. En effet, l'*ER process* tend à accélérer la divergence des séquences régulatrices. Bien qu'indirecte, la pression de sélection sur les séquences régulatrices est une pression positive : de nouvelles séquences régulatrices envahissent régulièrement la population. Ainsi, les séquences régulatrices proches des gènes tendent, à cause du *ER process*, à être remplacées plus souvent que si ces séquences évoluaient neutralement. Plus l'*ER process* est fort, plus vite les séquences sont censées diverger. On s'attend donc également à ce que les séquences régulatrices proches des gènes divergent davantage que les séquences régulatrices plus éloignées. Nous pouvons ainsi prédire un motif qui, à notre connaissance, ne peut être expliqué que par l'*ER process* : la divergence des séquences régulatrices proches du gène est plus grande que celle des séquences plus éloignées, et est supérieure à l'attendu neutre. En utilisant des données de divergence entre *Mus musculus* et *Rattus norvegicus* tirées de la banque de données Ensembl, nous avons pu trouver ce signal, faible car noyé dans un fond de sélection stabilisante, mais détectable. Il s'agit là de la première confirmation empirique concrète du processus que nous avons étudié. D'autres travaux seront nécessaires pour affiner ce signal, ou développer d'autres méthodes de vérification.

Bilan

Que faut-il retenir de cette thèse ?

En utilisant des modèles théoriques, nous avons découvert que l'avantage à cacher les mutations délétères peut entraîner chez les diploïdes des conséquences étonnantes et jusqu'à présent inconnues. Deux processus peuvent être induits par cet avantage. Lorsque les chromosomes homologues ne sont pas génétiquement isolés, on s'attend à ce que la force des promoteurs génétiques suffisamment proches des gènes augmente. En réponse, les autres régulateurs des gènes sont susceptibles d'évoluer pour maintenir des profils d'expression optimaux. Ce processus accélère notablement le remplacement des séquences régulatrices, et pourraient expliquer pourquoi celles-ci divergent si rapidement. Le second processus concerne le cas où les chromosomes homologues sont considérablement isolés génétiquement (pas ou peu de sexe, de recombinaison et de conversion génique). Dans ce cas, on prédit que l'expression des gènes devient progressivement haploïde. Une copie de chaque gène sur deux est progressivement éteinte, et dégénère. Dans le cas plus précis des chromosomes sexuels (les chromosomes X et Y sont génétiquement isolés), on remarque que ce processus devrait participer à l'extinction et à la dégénérescence du chromosome Y. Il pourrait ainsi permettre d'expliquer cette évolution si particulière des chromosomes sexuels, surtout en grandes populations, là où les théories actuelles sont limitées.

Ces processus sont remarquables en cela qu'ils nécessitent uniquement que les individus soient diploïdes. Ils ont pu donc être à l'œuvre depuis plus d'un milliard d'années, depuis que les eucaryotes sont apparus. Même si l'avantage sélectif de cacher des allèles délétères est faible, cet avantage est présent depuis si longtemps, dans une si large partie de l'arbre du vivant, qu'il ne peut être négligeable. Nous avons donc formulé un certain nombre de prédictions sur les conséquences que pouvaient avoir cet avantage sur les séquences régulatrices. Certaines d'entre elles sont testables. En collaborant avec Julien Dutheil et Gustavo Barroso de l'Institut Max Planck pour la Biologie Evolutive (Plön, Allemagne), nous avons pu directement vérifier une de nos prédictions.

A l'issu de cette thèse, tout un pan de l'évolution des séquences régulatrices semblent s'ouvrir. Il est urgent d'intensifier le travail de vérification des multiples prédictions que nous avons faites, et de quantifier l'impact des processus que nous avons décrits.

Annexe

Complexification de l'architecture des réseaux de régulation par l'*ER process*

Dans le Chapitre I, nous avons émis l'idée que l'*ER process*, en favorisant les mutations qui augmentent la force des promoteurs, pouvaient entraîner une complexification du réseau de régulation. Par exemple, une séquence régulatrice pourrait augmenter sa force en multipliant les séquences de fixation pour divers facteurs de transcription. Nous avons voulu, à travers la construction d'un nouveau modèle, préciser cette prédiction et étudier l'influence globale du *ER process* sur les régions non-codantes en amont des gènes.

Le Modèle

Notre modèle est constitué d'un locus gène, soumis à des mutations récurrentes délétères (avec un taux u_A). L'effet sélectif d'un allèle mutant est tiré dans une loi exponentielle négative de paramètre *1/s*. Le coefficient de dominance par défaut des mutations délétères (dominance lorsque les deux copies du gène sont exprimées en mêmes quantités) est fixé à $h = 0.25$. En amont de ce gène, une région non codante contrôle l'expression en *cis*. Elle contient des séquences de reconnaissance de facteurs de transcription, qu'on va nommer séquences promotrices ou promoteur. Au départ, un seul promoteur, de force initiale $e_0 = 10$ ($Log(e_0) = 1$), est présent au milieu de la zone régulatrice. De nouveaux promoteurs peuvent apparaître, soit *de novo* par mutation de séquences non codante (avec un taux u_{nov}), soit par duplication de promoteurs existant (avec un taux u_{dup}). Deux types de duplication sont possibles. Avec une probabilité égale à p_{LD}, la duplication se déroule à longue distance : la localisation du nouveau promoteur est tirée dans une loi normale, de moyenne égale à la localisation initiale du promoteur dupliqué et de variance σ_{LD}^2. Avec une probabilité égale à $1 - p_{LD}$, la duplication est à courte distance : la variance de la loi normale est égale à $\sigma_{SD}^2 < \sigma_{LD}^2$. Les promoteurs peuvent également être détruits par délétion (avec un taux u_{del}). Enfin, les promoteurs sont soumis à des mutations qui modifient leur force (avec un taux u_E). Ces mutations modifient la force des promoteurs de la même façon que dans nos autres modèles (*log*-force modifiée de façon additive via une loi normale centrée de variance σ_E^2).

Les individus sont soumis à deux forces sélectives. La première est la sélection purifiante sur le gène, qui élimine les mutations délétères. Si l'on note s_1 et s_2 les effets sélectifs des deux copies du gène d'un individu telles que $s_1 > s_2$, e_1 et e_2 les forces cumulées des régions régulatrices correspondantes, alors la fitness W_A de cet individu est calculée comme :

$$W_A = (1 - s_2) + (s_2 - s_1)\left(\frac{e_1}{e_1 + e_2}\right)^{-\frac{Log(h)}{Log(2)}} \quad (1)$$

Les individus peuvent être également soumis à une pression de sélection stabilisante sur les niveaux d'expression. La fitness des individus sur ce trait W_S est calculée, en notant I l'intensité de la sélection stabilisante et e_{opt} le niveau d'expression optimale, comme :

$$W_S = e^{-\left(Log(e_1)+Log(e_2)-2Log(e_{opt})\right)^2} \quad (2)$$

La fitness totale d'un individu, W, est finalement calculée en multipliant ces deux composantes de fitness : $W = W_A \; W_S$. A chaque génération, les parents d'un descendant sont tirés au hasard, et acceptés avec une probabilité égale à leur fitness. Un chromosome, éventuellement recombiné, est choisi chez chaque parent pour se transmettre au descendant.

Les promoteurs sont définis par leur force et par leur position dans la région non-codante. Plus un promoteur est situé loin en amont du gène, plus le taux de recombinaison avec le gène augmente. On considère qu'il y a, en moyenne, 0.1 événement de recombinaison par individu et par génération. La localisation des événements de recombinaison est tirée au hasard dans une loi continue.

On suit, au cours du temps (*300 000* générations), le nombre moyen de promoteurs par haplotype, la *log*-force moyenne des promoteurs, la *log*-force totale moyenne des régions régulatrices (définie comme l'addition des *log*-forces des promoteurs présents dans la région) ainsi que le gradient de *log*-force et le gradient de fréquence de présence des promoteurs le long de la région non-codante. Ces gradients sont calculés comme étant la pente des régressions linéaires de la *log*-force et de la fréquence des promoteurs en fonction de l'éloignement par rapport au gène. Des gradients négatifs signifient plus de promoteurs, et des promoteurs plus forts, à proximité du gène. Les simulations sont répétées 70 fois, et les résultats sont moyennés.

Les résultats présentés ici ont été obtenu pour : $u_A = 10^{-3}$, $u_E = 10^{-4}$, $u_{del} = 2x10^{-5}$, $u_{dup} = 10^{-5}$, $u_{nov} = 10^{-5}$, $p_{LD} = 0.2$, $\sigma_{SD} = 0.01$, $\sigma_{LD} = 0.2$, $\sigma_E = 0.1$. Les simulations ont été réalisées pour différentes valeurs de *s* et de *I*. Lorsque $I = 0$ et $s = 0$, aucune pression de sélection ne s'applique aux individus, les séquences évoluent neutralement. Lorsque $I = 0$ et $s = 0.2$, la sélection purifiante est en action, et l'*ER process* peut avoir lieu. L'escalade éventuelle des forces des promoteurs n'est pas contrecarrée par la sélection stabilisante. C'est en revanche le cas quand $s = 0,2$ et $I = 10^{-5}$ ou $I = 10^{-3}$. Enfin, quand $s = 0$ et $I = 10^{-5}$ ou $I = 10^{-3}$, seule la sélection stabilisante à un impact sur l'évolution des séquences régulatrices, tandis que le gène évolue neutralement.

Sur la Fig 2, on représente la *log*-force moyenne et la fréquence des promoteurs en fonction de leur éloignement au gène. Un éloignement de 0.1, par exemple, correspond à des promoteurs situés à des positions entre 0.05 et 0.1.

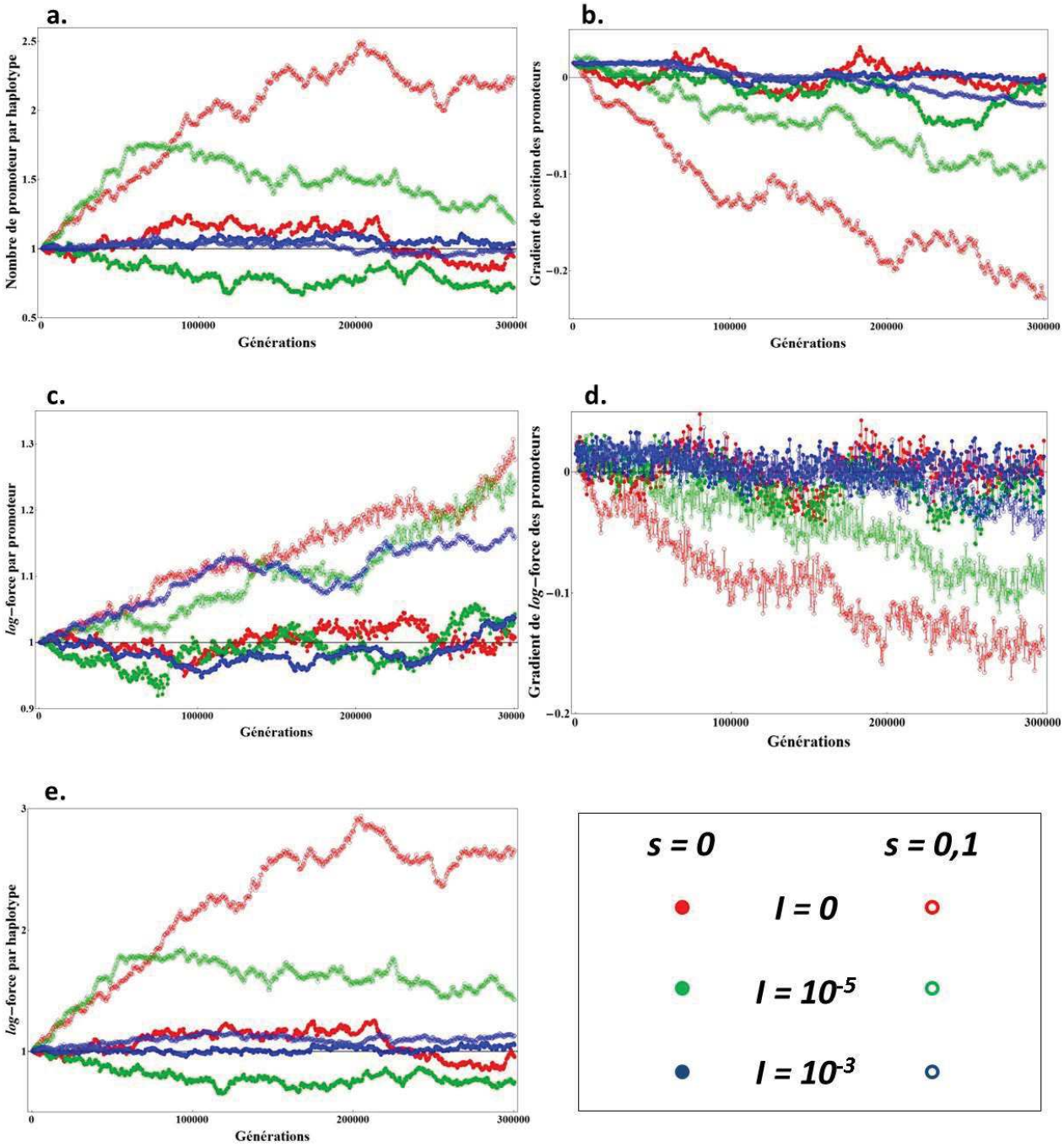


Figure 1. Evolution des régions non-codantes au cours du temps. On représente ici le nombre moyen de promoteurs par région non-codante (**a**), le gradient de fréquence des promoteurs le long des régions non-codantes (**b**), la *log*-force moyenne des promoteurs (**c**), le gradient de *log*-force des promoteurs le long des régions non-codantes (**d**) et la *log*-force moyenne cumulée des régions non-codantes (**e**). Les simulations sont réalisées pour $s = 0$ (cercles pleins) et $s = 0,2$ (cercles vides). Trois intensités de sélection stabilisante sont explorées : $I = 0$ (rouge), $I = 10^{-5}$ (vert) et $I = 10^{-3}$ (bleu).

Résultats

Sur la Fig 1, on représente l'évolution au cours du temps du nombre de promoteurs, de leur *log*-force, de la *log*-force des régions non-codantes entières et des gradients de *log*-force et de fréquence.

Dans les cas où l'*ER process* a lieu (*s = 0.2*), le nombre de promoteurs sur la région non-codante augmente. Cette augmentation est contrecarrée par la sélection stabilisante sur les niveaux d'expression : ele est moins forte à faible sélection stabilisante ($I = 10^{-5}$), et disparait à des intensités de sélection stabilisante supérieures *(*$I = 10^{-3}$*).* Ces promoteurs, quand ils sont soumis au *ER process*, tendent à "migrer" plus près du gène que leur position initiale, comme le révèlent des gradients de fréquence en baisse et négatifs quand *s = 0.2*. Ainsi, on voit clairement sur la Fig 2.a que les promoteurs sont plus nombreux proche du gène, et que leur fréquence diminue à mesure que l'on s'éloigne du gène.

L'*ER process* a également pour conséquence d'augmenter la force moyenne des promoteurs. La sélection stabilisante sur les niveaux d'expression tend à ralentir cette augmentation, mais elle est toujours importante à $I = 10^{-3}$. Dans l'ensemble, l'*ER process* tend à augmenter le nombre de promoteurs, et à augmenter leur force moyenne. L'*ER process* tend donc à augmenter la force globale des régions non-codantes, et donc les niveaux d'expression. Cette augmentation semble linéaire quand il n'y a pas de sélection stabilisante sur les niveaux d'expression (*I = 0*). La sélection stabilisante tend à limiter cette augmentation. La force globale des régions non-codantes, et donc les niveaux d'expression, est le résultat d'un compromis entre *ER process*, qui tire les niveaux d'expression vers le haut, et sélection stabilisante, qui tend à les ramener vers l'optimum.

Enfin, on observe que les gradients de *log*-force tendent à diminuer et à être négatif quand l'*ER process* est en action, et ce d'autant plus que la sélection stabilisante est faible. Lorsqu'il n'y a pas de sélection stabilisante, les promoteurs proches du gène augmentent davantage que ceux loin du gène parce qu'ils recombinent moins souvent avec celui-ci. Lorsque la sélection stabilisante sur les niveaux d'expression est présente, à cela s'ajoute la nécessité, pour les promoteurs loin du gène, de compenser l'augmentation de force des promoteurs proches du gène pour assurer des niveaux d'expression plus proches de l'optimum (voir Chapitre II). Ainsi, on peut observer sur la Fig 2.b que les promoteurs proches du gène

tendent à augmenter en force, tandis que les promoteurs plus éloignés tendent à diminuer en force.

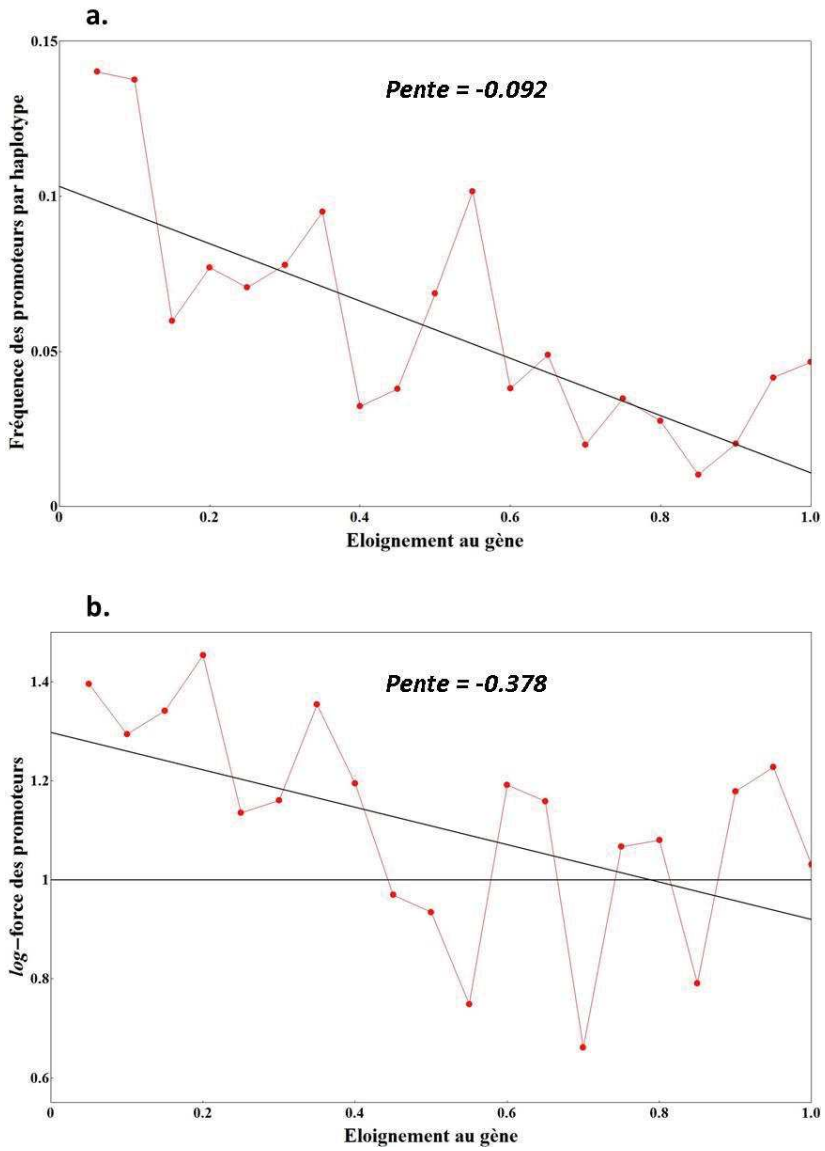


Figure 2. Distribution de la fréquence (**a**) et de la *log*-force moyenne (**b**) des promoteurs le long des régions non-codantes après 300 000 générations. Les populations simulées ici sont soumises à la sélection purifiante (*s = 0,2*) et à une sélection stabilisante faible sur les niveaux d'expression ($I = 10^{-5}$).

Discussion

Ces résultats confirment que l'*ER process* est susceptible de durablement influencer l'architecture des réseaux de régulation. Tout d'abord, ce processus tend à complexifier le réseau, en multipliant les séquences promotrices. La sélection stabilisante tend à limiter cette multiplication des promoteurs car celle-ci entraîne, au moins transitoirement, des niveaux d'expression sous-optimaux. Ensuite, on constate que l'*ER process* tend à favoriser une certaine organisation des promoteurs en amont du gène. Les promoteurs les plus forts tendent à se concentrer près du gène, tandis que des promoteurs moins forts, voir inhibiteurs, sont plutôt présents plus loin du gène.

La pression de sélection à l'origine de l'*ER process* est faible, comme indiqué dans le Chapitre I. Il faut ainsi plusieurs milliers, voire dizaines de milliers de génération pour voir cette organisation apparaître. Cependant, les eucaryotes sont apparus il y a plus d'un milliard d'années, ce qui a laissé amplement le temps à l'*ER process* d'influencer l'organisation des réseaux de régulation.

Les résultats de ce modèle sont des résultats préliminaires. Des simulations doivent encore être réalisées afin de diminuer la dérive génétique (en augmentant la taille des populations), car les simulations restent pour le moment très bruitées. Il est également nécessaire d'observer l'évolution du réseau sur un temps plus long (en augmentant le nombre de génération) et de mieux appréhender l'influence de la sélection stabilisante (simulations avec d'autres intensités de sélection stabilisante). Néanmoins, ces résultats sont encourageants, et amènent à penser que plusieurs caractéristiques des réseaux d'expression, notamment leur complexité (en termes de séquences promotrices), sont fortement influencées par l'*ER process*. Cette théorie est assez différentes des théories visant à expliquer l'évolution de la forte complexité des systèmes génétiques des eucaryotes (en comparaison aux procaryotes) par des processus neutres dus à leur plus petite taille de population (LYNCH 2007). Ici, l'existence de réseaux complexes de régulation chez les eucaryotes dériverait de la diploïdie (qui génère automatiquement la compétition pour l'expression entre *cis*-régulateurs et donc l'*ER process*) plutôt que de leur taille de population supposée plus faible que celle des procaryotes.

Bibliographie

ABZHANOV A., PROTAS M., GRANT B. R., GRANT P. R., TABIN C. J., 2004 Bmp4 and morphological variation of beaks in Darwin's finches. Science **305**: 1462–5.

ALON U., 2007 Network motifs: theory and experimental approaches. Nat. Rev. Genet. **8**: 450–461.

BIGGIN M. D., 2014 Animal Transcription Networks as Highly Connected, Quantitative Continua. Dev. Cell **21**: 611–626.

BONIFER C., COCKERILL P. N., 2011 Chromatin Mechanisms Regulating Gene Expression in Health and Disease BT - Epigenetic Contributions in Autoimmune Disease. In: Ballestar E, *Madame Curie Bioscience Database*, Springer US, pp. 12–25.

BRANDMAN O., MEYER T., 2008 Feedback Loops Shape Cellular Signals in Space and Time. Science **322**: 390–395.

BURT A., TRIVERS R., 2009 *Genes in conflict: the biology of selfish genetic elements*. Harvard University Press.

BUTLER J. E. F., KADONAGA J. T., 2002 The RNA polymerase II core promoter: a key component in the regulation of gene expression. Genes Dev. **16**: 2583–2592.

CAILLEAU A., CHEPTOU P.-O., LENORMAND T., 2010 Ploidy and the evolution of endosperm of flowering plants. Genetics **184**: 439–453.

CHARLESWORTH B., 1979 Evidence against Fisher's theory of dominance. Nature **278**: 848–849.

COOPER T. F., 2003 Parallel changes in gene expression after 20,000 generations of evolution in *Escherichia coli*. Proc. Natl. Acad. Sci. U. S. A. **100**: 1072–1077.

CROMBACH A., HOGEWEG P., 2008 Evolution of Evolvability in Gene Regulatory Networks. PLoS Comput. Biol. **4**: e1000112.

DENVER D. R., MORRIS K., STREELMAN J. T., KIM S. K., LYNCH M., THOMAS W. K., 2005 The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*. Nat. Genet. **37**: 544–8.

ERCAN S., 2015 Mechanisms of X chromosome dosage compensation. J. Genomics **3**: 1–19.

ESTELLER M., 2007 Epigenetic gene silencing in cancer: the DNA hypermethylome. Hum. Mol. Genet. **16**: R50–R59.

FEREA T. L., BOTSTEIN D., BROWN P. O., ROSENZWEIG R. F., 1999 Systematic changes in gene expression patterns following adaptive evolution in yeast. Proc. Natl. Acad. Sci. U. S. A. **96**: 9721–9726.

FISHER R. A., 1928 The Possible Modification of the Response of the Wild Type to Recurrent Mutations. Am. Nat. **62**: 115–126.

GILAD Y., OSHLACK A., SMYTH G. K., SPEED T. P., WHITE K. P., 2006 Expression profiling in primates reveals a rapid evolution of human transcription factors. Nature **440**: 242–5.

HE F., ZHANG X., HU J., TURCK F., DONG X., GOEBEL U., BOREVITZ J., MEAUX J. DE, 2012 Genome-wide Analysis of Cis-regulatory Divergence between Species in the *Arabidopsis* Genus. Mol. Biol. Evol. **29**: 3385–3395.

HURST L. D., RANDERSON J. P., 2000 Dosage, Deletions and Dominance: Simple Models of the Evolution of Gene Expression. J. Theor. Biol. **205**: 641–647.

JENKINS D. J., STEKEL D. J., 2010 *De Novo* Evolution of Complex, Global and Hierarchical Gene Regulatory Mechanisms. J. Mol. Evol. **71**: 128–140.

KACSER H., BURNS J. A., 1981 The Molecular Basis of Dominance. Genetics **97**: 639 LP – 666.

KHAITOVICH P., WEISS G., LACHMANN M., HELLMANN I., ENARD W., MUETZEL B., WIRKNER U., ANSORGE W., PÄÄBO S., 2004 A Neutral Model of Transcriptome Evolution. PLoS Biol. **2**: e132.

KING M.-C., WILSON A., 1975 Evolution at two levels in humans and chimpanzees. Science **188**: 107–116.

LEE T. I., YOUNG R. A., 2000 Transcription of Eukaryotic Protein-Coding Genes. Annu. Rev. Genet. **34**: 77–137.

LEMOS B., MEIKLEJOHN C. D., CÁCERES M., HARTL D. L., 2005 Rates of divergence in gene expression profiles of primates, mice, and flies: stabilizing selection and variability

among functional categories. Evolution. **59**: 126–37.

LENSKI R. E., OFRIA C., PENNOCK R. T., ADAMI C., 2003 The evolutionary origin of complex features. Nature **423**: 139–144.

LO H. S., WANG Z., HU Y., YANG H. H., GERE S., BUETOW K. H., LEE M. P., 2003 Allelic Variation in Gene Expression Is Common in the Human Genome. Genome Res. **13**: 1855–1862.

LUDWIG M. Z., BERGMAN C., PATEL N. H., KREITMAN M., 2000 Evidence for stabilizing selection in a eukaryotic enhancer element. Nature **403**: 564–7.

LYNCH M., 2007 The frailty of adaptive hypotheses for the origins of organismal complexity. Proc. Natl. Acad. Sci. **104**: 8597–8604.

MANNA F., MARTIN G., LENORMAND T., 2011 Fitness Landscapes: An Alternative Theory for the Dominance of Mutation. Genetics **189**: 923–937.

METZGER B. P. H., DUVEAU F., YUAN D. C., TRYBAN S., YANG B., WITTKOPP P. J., 2016 Contrasting frequencies and effects of cis- and trans-regulatory mutations affecting gene expression. Mol. Biol. Evol. .

MILO R., SHEN-ORR S., ITZKOVITZ S., KASHTAN N., CHKLOVSKII D., ALON U., 2002 Network Motifs: Simple Building Blocks of Complex Networks. Science (80-.). **298**: 824 – 827.

MOREY C., AVNER P., 2010 Genetics and epigenetics of the X chromosome. Ann. N. Y. Acad. Sci. **1214**: E18–E33.

OGBOURNE S., ANTALIS T. M., 1998 Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochem. J. **331**: 1–14.

ORR H. A., 1991 A test of Fisher's theory of dominance. Proc. Natl. Acad. Sci. **88** : 11413–11415.

ORR H. A., KIM Y., 1998 An Adaptive Hypothesis for the Evolution of the Y Chromosome. Genetics **150**: 1693–1698.

OTTO S. P., GOLDSTEIN D. B., 1992 Recombination and the evolution of diploidy. Genetics **131**: 745–751.

OTTO S. P., YONG P., 2002 The evolution of gene duplicates. Adv. Genet. **46**: 451–483.

PROULX S. R., PHILLIPS P. C., 2005 The Opportunity for Canalization and the Evolution of Genetic Networks. Am. Nat. **165**: 147–162.

RAYMOND M., CHEVILLON C., GUILLEMAUD T., LENORMAND T., PASTEUR N., 1998 An overview of the evolution of overproduced esterases in the mosquito *Culex pipiens*. Philos. Trans. R. Soc. B Biol. Sci. **353**: 1707–1711.

SEN R., GROSSCHEDL R., 2010 Memories of lost enhancers. Genes Dev. **24**: 973–979.

SERFLING E., JASIN M., SCHAFFNER W., 1985 Enhancers and eukaryotic gene transcription. Trends Genet. **1**: 224–230.

SOYER O. S., BONHOEFFER S., 2006 Evolution of complexity in signaling pathways. Proc. Natl. Acad. Sci. **103**: 16337–16342.

STEIGE K. A., REIMEGÅRD J., KOENIG D., SCOFIELD D. G., SLOTTE T., 2015 Cis-Regulatory Changes Associated with a Recent Mating System Shift and Floral Adaptation in *Capsella*. Mol. Biol. Evol. **32**: 2501–2514.

STERN D. L., 1998 A role of Ultrabithorax in morphological differences between *Drosophila* species. Nature **396**: 463–466.

SZAFRANIEC K., WLOCH D. M., SLIWA P., BORTS R. H., KORONA R., 2003 Small fitness effects and weak genetic interactions between deleterious mutations in heterozygous loci of the yeast *Saccharomyces cerevisiae*. Genet. Res. **82**: 19–31.

WAGNER G. P., BÜRGER R., 1985 On the evolution of dominance modifiers II: a non-equilibrium approach to the evolution of genetic systems. J. Theor. Biol. **113**: 475–500.

WALTERS M. C., FIERING S., EIDEMILLER J., MAGIS W., GROUDINE M., MARTIN D. I., 1995 Enhancers increase the probability but not the level of gene expression. Proc. Natl. Acad. Sci. **92**: 7125–7129.

WEIRAUCH M. T., HUGHES T. R., 2010 Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends Genet. **26**: 66–74.

WHITEHEAD A., CRAWFORD D. L., 2006a Variation within and among species in gene expression: raw material for evolution. Mol. Ecol. **15**: 1197–1211.

WHITEHEAD A., CRAWFORD D. L., 2006b Neutral and adaptive variation in gene expression. Proc. Natl. Acad. Sci. **103**: 5425–5430.

WITTKOPP P. J., KALAY G., 2012 Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nat. Rev. Genet. **13**: 59–69.

WITTKOPP P. J., VACCARO K., CARROLL S. B., 2002 Evolution of yellow gene regulation and pigmentation in *Drosophila*. Curr. Biol. **12**: 1547–56.

WRAY G. A., 2007 The evolutionary significance of cis-regulatory mutations. Nat. Rev. Genet. **8**: 206–216.

WRAY G. A., HAHN M. W., ABOUHEIF E., BALHOFF J. P., PIZER M., ROCKMAN M. V, ROMANO L. A., 2003 The evolution of transcriptional regulation in eukaryotes. Mol. Biol. Evol. **20**: 1377–419.

WRIGHT S., 1934a Fisher's Theory of Dominance. Am. Nat. **63**: 274–279.

WRIGHT S., 1934b Physiological and Evolutionary Theories of Dominance. Am. Nat. **68**: 24–53.

Remerciements

Tout d'abord, merci à ma famille qui m'a poussé, comme il se doit, à bien travailler à l'école. Grâce à elle, j'ai pu intégrer l'Ecole Normale Supérieure de Paris et obtenir une bourse de thèse beaucoup plus facilement que toutes les personnes qui passent par l'université. A présent, je suis également docteur, et c'est à elle que je le dois.

Un grand merci également à Thomas pour m'avoir donné l'opportunité de travailler sur ce projet et pour m'avoir accompagné pendant 4 ans. Ses compétences, et la finesse de ses analyses m'ont été d'une aide précieuse pour mener à bien ces travaux. Je me rappellerai longtemps nos longues discussions pour interpréter nos résultats, discussions qui, il faut le dire, ont grandement contribué aux ventes de Doliprane. Merci aussi de m'avoir emmené en séminaire / colloque à Aussois (Savoie) en 2013 et à Porto Rico en 2014 : il y a pire comme endroit pour partir en vaconf', euh non pardon, pour aller participer à des échanges scientifiques intéressants et novateurs. Merci enfin pour ses conseils quant à mon avenir. Je ne sais pas si je les suivrai, mais ça fait toujours plaisir d'avoir un superviseur qui s'en préoccupe.

Merci aux membres de mes comités de thèse : Pierre Boursot, Sylvain Glémin, Bernard de Massy, Christoph Haag et Gabriel Marais. Leurs compétences diverses m'ont été très utiles à certains moments. Mes travaux sont à la jonction de nombreuses questions évolutives (régulation de l'expression, conflits génétiques, recombinaison, modes de reproduction, chromosomes sexuels), et il a pu m'arriver de me perdre dans la bibliographie de tant de domaines. Ils ont su m'aiguiller sur certains points critiques, et je les en remercie. Merci aussi aux membres du jury d'avoir accepté de prendre du temps pour examiner mes travaux de thèse : j'espère qu'ils vous ont intéressé.

Bien sûr, un grand merci à la technique ! Merci à Marie-Claude Quidoz, qui gère la plate-forme de calcul du CEFE, que j'ai beaucoup utilisé. Son aide a été précieuse quand les bugs passaient à l'offensive, et son humour pince-sans-rire est toujours revigorant. Merci également à la plate-forme de calcul MBB, sur laquelle j'ai effectué de nombreux calculs, et à ceux qui la gèrent.

Je remercie l'ensemble des personnes qui travaillent au CEFE, merci d'en faire tous les jours un lieu serein. Malgré tout ce que j'ai pu râler ou contester au CEFE, la gentillesse de ces membres n'est plus à démontrer. Merci aux copains et copines qui se reconnaîtront, post-

docs, thésards, stagiaires : on aura quand même bien rigolé ! Dédicace toute particulière à Noémie Harmand et Eva Lievens, mes sœurs de thèse (bros for life !), et à Elsa Noël, Pascal Marrot et Tim Janicke, avec qui j'ai partagé mon bureau.

Enfin, un grand, un immense merci à la ZAD (Zone d'Activité Démocratique) de Las Rébès. J'étais perdu, mais grâce à cet endroit merveilleux, la lumière s'est rallumée. Grâce aux copains et aux copines, je me suis rappelé que la vie valait la peine de se battre. Je ne peux pas les nommer, mais les personnes concernées se reconnaîtront. Sans vous, ces derniers mois auraient été un calvaire, grâce à vous, ç'a été parmi les plus belles semaines de ma vie. La ZAD est finie, mais je ne suis pas prêt de vous oublier.

Plus généralement, merci à tous les militant-es du printemps 2016 : ça fait plaisir de voir que je ne suis pas le seul à ne pas avoir abandonné l'avenir aux banquiers et autres industriels de tout poil. Merci à vous de recréer l'espoir d'un avenir au milieu du désastre moderne. Merci, car à quoi sert-il d'accumuler des connaissances si le futur est mort ?

COMPETITION POUR LA TRANSCRIPTION ET EVOLUTION DE L'EXPRESSION GENETIQUE CHEZ LES DIPLOÎDES

Les séquences non-codantes régulatrices de l'expression des gènes sont tout aussi importantes pour le phénotype d'un individu que les séquences codantes. De nombreux travaux se sont attachés à identifier les forces influençant l'évolution de ces séquences non-codantes. Ici, nous théorisons une nouvelle force sélective influençant potentiellement l'évolution de certaines séquences régulatrices. En utilisant des modèles multi-locus, nous montrons que les promoteurs génétiques les plus forts (activateurs de la transcription) gagnent un avantage à voir la copie du gène qui leur est associée (située sur le même chromosome) davantage exprimée que la copie homologue, contrôlée par un promoteur homologue moins fort. La surexpression des copies associées aboutit à une meilleure purge des mutations délétères chez ces copies, et ainsi à une association génétique entre promoteurs forts et contexte génétique favorable. Si la recombinaison entre le gène et le promoteur est suffisamment faible pour que cette association persiste, la force des promoteurs est sélectionnée pour augmenter. L'escalade des forces des promoteurs ne conduit pas forcément à une surproduction de protéines : d'autres régulateurs peuvent co-évoluer pour maintenir un niveau d'expression optimal, à condition que la sélection stabilisante tolère des niveaux d'expression transitoirement sub-optimaux. En variant les modes de reproduction, nous avons montré que ce nouveau processus sélectif ne menait pas nécessairement à une escalade de la force des promoteurs. Lorsque les chromosomes sont suffisamment isolés génétiquement (peu de recombinaison, peu de fécondation croisée), la sélection pour des associations génétiques favorables mène à une divergence des chromosomes : un chromosome accumule des promoteurs forts et possède des copies viables du gène, tandis que le chromosome homologue accumule des promoteurs faibles et des mutations délétères sur le gène. Dans le cas de lignées clonales peu ou pas recombinantes, on s'attend ainsi à observer une haploïdisation de l'expression des gènes : une copie de chaque gène concerné est éteinte et dégénère. Cette divergence s'applique aussi à des chromosomes sexuels ayant cessé de recombiner : on a pu montrer que la divergence des chromosomes menait à une extinction et une dégénérescence des gènes situés sur les chromosomes Y, et à une surexpression des gènes situés sur le chromosome X. En utilisant notre modèle, on propose ainsi une nouvelle théorie pour expliquer l'évolution des chromosomes sexuels non-recombinants. Enfin, on a utilisé des données de divergence entre *Mus musculus et Rattus norvegicus* pour isoler un signal ne pouvant être expliqué que par une sélection positive pour des promoteurs proximaux plus forts. Ce signal est faible, mais détectable, nous permettant d'apporter une première confirmation empirique du processus d'escalade des forces des promoteurs.

Mots-clés : modèle multi-locus, évolution de l'expression, sélection, dominance

---

COMPETITION FOR TRANSCRIPTION AND GENE EXPRESSION EVOLUTION IN DIPLOIDS

Non-coding sequences, that regulate gene expression, are as important as coding sequences to determine phenotypes. Many studies have identified the main forces affecting regulatory sequence evolution. Here, we theoretically identify a new selective force that may also play a role in this matter. Using multi-locus models, we show that stronger (activating more transcription) enhancers gain some benefit in having its associated gene copy more expressed than the homolog gene copy, controlled by a weaker homolog enhancer. Overexpressed gene copies are better purged from deleterious mutations, such that stronger enhancers get associated with a better genetic background. If recombination between the gene and the enhancer is low enough for this association to persist, enhancer strength selectively increases. Enhancer strength escalation does not necessarily lead to protein overproduction. Other regulators may indeed co-evolve to maintain optimal expression levels, provided that stabilizing selection allows for transitory sub-optimal expression levels. Implementing in the models different reproductive systems, we show that this new selective process does not necessarily lead to an enhancer strength escalation. When chromosomes are genetically isolated enough (little recombination, little outcrossing, selection for favorable genetic associations leads to chromosome divergence: one accumulates stronger enhancers and viable gene alleles, while the other accumulates weaker enhancers and deleterious gene mutations. For non-recombining clonal lineages, we expect gene expression to become haploid: for each gene, one copy is shut down and degenerates. Such divergence also applies to non-recombining sex chromosomes. We show that in such case, chromosome divergence leads to a shut down and degeneration of Y chromosome genes, and to an overexpression of genes located on X chromosomes. With our model, we propose a new theory to explain sex chromosome evolution after they stop recombining. Finally, we used divergence data between *Mus musculus* and *Rattus norvegicus* to find a signal that can only be explained by positive selection for stronger proximal enhancers. This signal is weak, but significant: this is the first empirical confirmation of enhancer strength escalation process we studied here.

Key-words: multi-locus models, expression evolution, selection, dominance