The Y rescued by the X ? : evolution of dosage compensation in humans and other questions on sex chromosome evolution in eukaryotes

Eugénie Pessia

N° 261-2013 Année 2013

Thèse de l'université de lyon

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON 1

pour l'obtention

du Diplôme de doctorat
(arrêté du 7 août 2006)

soutenue publiquement le
12 décembre 2013

par

Eugénie Pessia

Comment le X vient-il à la rescousse du Y ? Évolution de la compensation de dosage des XY humains et autres questions sur l'évolution des chromosomes sexuels eucaryotes

Directeur de thèse : Gabriel Marais

Jury :

Tatiana Giraud	Rapporteur
Judith Mank	Rapporteur
Gabriel Marais	Directeur de thèse
Marie Sémon	Examinateur
Frédéric Veyrunes	Examinateur
Cristina Vieira	Examinateur

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. François-Noël GILLY**
Vice-président du Conseil d'Administration	M. le Professeur Hamda BEN HADID
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur Philippe LALLE
Vice-président du Conseil Scientifique	M. le Professeur Germain GILLET
Directeur Général des Services	M. Alain HELLEU

COMPOSANTES SANTE

Faculté de Médecine Lyon Est – Claude Bernard	Directeur : M. le Professeur J. ETIENNE
Faculté de Médecine et de Maïeutique Lyon Sud – Charles Mérieux	Directeur : Mme la Professeure C. BURILLON
Faculté d'Odontologie	Directeur : M. le Professeur D. BOURGEOIS
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : Mme la Professeure C. VINCIGUERRA
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. MATILLON
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. DE MARCHI
Département Biologie	Directeur : M. le Professeur F. FLEURY
Département Chimie Biochimie	Directeur : Mme le Professeur H. PARROT
Département GEP	Directeur : M. N. SIAUVE
Département Informatique	Directeur : M. le Professeur S. AKKOUCHE
Département Mathématiques	Directeur : M. le Professeur A. GOLDMAN
Département Mécanique	Directeur : M. le Professeur H. BEN HADID
Département Physique	Directeur : Mme S. FLECK
Département Sciences de la Terre	Directeur : Mme la Professeure I. DANIEL
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. COLLIGNON
Observatoire des Sciences de l'Univers de Lyon	Directeur : M. B. GUIDERDONI
Polytech Lyon	Directeur : M. P. FOURNIER
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. PIGNAULT
Institut Universitaire de Technologie de Lyon 1	Directeur : M. C. VITON
Institut Universitaire de Formation des Maîtres	Directeur : M. A. MOUGNIOTTE
Institut de Science Financière et d'Assurances	Administrateur provisoire : M. N. LEBOISNE

Résumé

L'évolution des chromosomes sexuels constitue un sujet d'étude important depuis près de 100 ans. Chez l'Homme, il a pu être démontré que les chromosomes X et Y étaient à l'origine une paire d'autosomes. Après l'acquisition du gène du déterminisme sexuel *Sry*, les chromosomes X et Y ont arrêté de recombiner sur une partie de leur longueur, ce qui a rendu l'évolution de leurs séquences indépendante. Ils se sont alors différenciés en deux chromosomes très différents. Le Y étant toujours seul dans le génome, il s'est arrêté de recombiner depuis l'arrêt de recombinaison. La sélection a alors dû agir sur l'ensemble de ses gènes comme un bloc ce qui l'a rendu beaucoup moins efficace par effets dits "Hill-Robertson". Ceci a causé entre autres une grande perte de gènes et une accumulation d'éléments transposables. Le Y a donc dégénéré de plus en plus au cours du temps, ce qui a mené certains auteurs à penser qu'il était amené à disparaître. Le X, au contraire, a pu conserver une sélection efficace puisqu'il recombine chez les femelles (XX) deux-tiers de son temps. Il a alors été considéré comme le chromosome sexuel immuable, qui n'aurait pas changé depuis qu'il était encore un autosome.

Cette vision dualiste des chromosomes sexuels n'est cependant pas tout à fait vraie. En effet, la dégénérescence du Y n'est pas un processus linéaire dans le temps : certains gènes sont très bien conservés même dans de "vieux" chromosomes (~180 millions d'années chez les mammifères). Le X, quant à lui, s'est adapté à la grande perte de gènes du Y, en mettant en place une compensation de dosage pour éviter que les mâles ne subissent des effets délétères suite à la diminution de leur expression des gènes X. Certains auteurs ont aussi supposé qu'une forme de recombinaison qui persiste entre les gènes homologues du X et du Y, la conversion génique X-Y, pourrait contribuer à limiter la dégénérescence du Y. Le X peut donc être vu comme un "sauveur", venant à la rescousse de son compagnon le chromosome Y dans ses difficultés liées à leur arrêt de recombinaison (du moins par crossing-overs).

Un premier pan de ma thèse concerne ces deux différents mécanimes de sauvetage du Y par le X. Premièrement, j'ai participé à une controverse sur la compensation de dosage chez les mammifères. Une hypothèse avait été proposée dans les années 60 par Susumo Ohno, proposant un mécanisme de compensation en deux temps. Chez les mâles, la perte de nombreux gènes sur le Y entraîne un déséquilibre de dosage car ces gènes qui étaient précédemment présents en deux copies sont devenus unicopie, soit une division d'expression par deux. Selon l'hypothèse d'Ohno, chez les mammifères en réponse à cela le X aurait doublé son expression, mais dans les deux sexes menant ainsi à une expression trop élevée chez les femelles. Ce deuxième problème de dosage aurait alors été résolu par la mise en place d'une inactivation aléatoire de l'un des deux X chez les femelles. Tandis que la deuxième partie de l'hypothèse d'Ohno, l'inactivation du X, a été très étudiée,

la première partie est restée spéculative jusqu'aux années 2000. En étudiant des données d'expression du X humain j'ai pu montrer, de manière concomitante avec d'autres auteurs, que la première partie de l'hypothèse d'Ohno n'est pas totalement vraie car seule une partie des gènes sont sur-exprimés. J'ai ensuite participé à l'écriture d'une revue visant à donner une explication alternative à la compensation de dosage pour l'évolution de l'inactivation du X chez les femelles mammifères. Deuxièmement, j'ai étudié la présence de conversion génique X-Y dans plusieurs gènes, au sein de nombreuses espèces de primates. Mes travaux me mènent à discuter le fait que ce type d'évènement soit effectivement favorisé par la sélection. Je pose l'hypothèse que ces conversions géniques ont été maintenues de manière neutre. Ces deux travaux ne vont pas dans le sens d'un chromosome X sauvant le Y avec beaucoup de zèle.

Dans un dernier temps, m'éloignant des espèces modèles, j'ai étudié les chromosomes sexuels particuliers d'une algue brune : *Ectocarpus siliculosus*. Cela m'a permis de vérifier si le scénario évolutif actuel des chromosomes sexuels est toujours valide dans un groupe d'eucaryotes séparé des animaux depuis plus d'un milliard d'années.

Mots-clés : évolution moléculaire, chromosomes sexuels, mammifères, algues brunes.

Abstract

Sex chromosome evolution has been an important subject of study for almost 100 years. In humans, it has been proved that the X and Y chromosomes were initially a pair of autosomes. After the acquisition of the sex-determining gene *Sry*, the X and Y stopped recombining on part of their length, their sequence evolution thus becoming independent. They then differentiated into two very different chromosomes. The Y, which is always alone in the genome, stopped recombining after this recombination arrest with the X. Selection then acted on all of its genes as a block, which made it a lot less efficient by "Hill-Robertson" effects. This caused a high gene loss and tranposable elements accumulation, among others. The Y thus became more and more degenerated along time, which led some authors to think the Y was doomed to disappear. The X, on the contrary, maintained an efficient selection as it recombines in females (XX) two-third of its time. This chromosome has thus been considered as the stable sex chromosome, which has not changed since when it was still an autosome.

This dualistic vision of sex chromosomes is however not exactly true. The Y degeneration is indeed not a linear process through time: some genes are well conserved even in "old" chromosomes (~180 million years in mammals). The X, meanwhile, adapted to the Y high gene loss by establishing a dosage compensation mechanism, in order to avoid that males suffer from the deleterious effects due to their X-linked genes decrease in expression. Some authors also suggested that X-Y gene conversion, a type of recombination persisting between X and Y homologous genes, could help limiting the Y degeneration. The X chromosome can thus be seen as a "savior" helping his fellow the Y in its difficulties caused by their recombination arrest (by crossing-overs at least).

The first part of my thesis concerns these two different mechanisms of the Y being rescued by the X. Firstly, I contributed to a controversy on mammalian dosage compensation. During the 60s Susumo Ohno hypothesized a two-step dosage compensation mechanism. In males, the high loss of Y-linked genes led to a dosage imbalance: these genes were previously present in two allelic copies and became unicopy, meaning that their expression has been halved. According to Ohno's hypothesis, in response to this imbalance the mammalian X would have doubled its expression in the two sexes, resulting in a to high expression in females. This second dosage imbalance would have been resolved by the random inactivation of one of the two Xs in females. Whereas the second part of Ohno's hypothesis, the X-chromosome inactivation, has been well studied, the first part remained speculative until the 2000s. I studied human X-linked expression data and was able to show, concomitantly with other authors, that the first part of Ohno's hypothesis is not totally true as only some of the X-linked genes are hyperexpressed. I later participated in the writing of a review aiming to give an alternative hypothesis for

the evolution of X-chromosome inactivation in mammalian females than dosage compensation. Secondly, I studied signatures of X-Y gene conversion in several genes within numerous primate species. My results led me to discuss if these events were indeed selected for. I hypothesize that these gene conversion events occurred in a neutral manner. These two different studies suggest that the X chromosome may not be as much a help for the Y as has been suggested.

Lastly, moving away from model species, I studied the peculiar sex chromosomes of a brown alga: *Ectocarpus siliculosus*. This work allowed me to test if the current hypotheses on sex chromosome evolution still hold in a eukaryotic group that diverged from animals more than one billion years ago.

Keywords: molecular evolution, sex chromosomes, mammals, brown algae.

Remerciements

Exercice ô combien classique que de remercier son entourage pour la réalisation d'un manuscrit à l'aspect pompeux, que ces mêmes personnes ne liront très probablement pas !
... Et pourtant, ce n'est pas pour faire comme tout le monde que j'écris ces lignes. Je tiens réellement, du plus profond de mon cœur, à remercier chaleureusement :

ma mère Jacqueline, pour avoir eu souvent plus d'ambition concernant ma réussite que moi-même

mon beau-père Alain et mon frère Mathieu, pour avoir essayé de me faire croire qu'ils comprenaient mon sujet de thèse

Madame Courrejou qui m'a enseigné la biologie en 1ère S, Messieurs Anselme et Aubert et Madame Ginestet qui m'ont enseigné les sciences en classes préparatoires, pour m'avoir fait découvrir la beauté des sciences et des découvertes scientifiques

Gabriel Marais, qui a été un directeur de thèse hors du commun sachant subtilement doser conseils, écoute et efficacité

ceux qui ont été d'un grand soutien et ont su être patients face à mes fameux râlages et autres coups de gueule (par ordre chronologique) : Johanna Crambes, Jafu, Myriam Roudy, Onis, Anthony Lossmann, Nicolas Estibals, Sophie Padié, Thomas Bigot, Marc Bailly-Bechet, Michaël Faubladier, Clothilde Deschamps, Julien Dutel, Sophia Ahmed et Laurent Modolo

ceux qui ont contribué à la bonne ambiance des essentielles pauses post-prandiales : Florent Lassalle, Bérénice Batut, Rémi Planel, Héloïse Philippon, Magali Semeria, Aline Muyle, Simon Penel, Yann Lesecque, Sylvain Mousset, Jos Käfer, Nicolas Rochette, Erika Kvikstad, Joanna Parmley, Dominique Guyot, Matthieu Barba

les collègues pleins de bons conseils et toujours disponibles : Laurent Duret, Marc Bailly-Bechet, Raquel Tavares, Sylvain Mousset, Vincent Daubin, Laurent Gueguen, Marie Sémon, Manolo Gouy, Bastien Bousseau, Daniel Kahn, Jean Thioulouse, Dominique Mouchiroud, Stéphane Delmotte, Bruno Spataro, Lionel Humblot

mes collaboratrices : Aoife McLysaght, Susana Coelho et Brigitte Crouau-Roy

et enfin : mes bras, pour avoir toujours été de mon côté ; mes jambes, pour m'avoir toujours supportée ; et mes doigts, sur lesquels je peux toujours compter

J'espère n'oublier personne, mais en réalité je ne suis pas dupe et je sais déjà que c'est faux !

Dédié aux nombreux individus (humains comme animaux) qui,
au cours de ma vie parfois mouvementée,
ont su m'apporter un rayon de soleil et de l'apaisement.
Sans vous, je ne serais pas la même femme,
et je ne serais probablement pas là où je suis aujourd'hui.

Contents

Introduction

Diversity of sex determination mechanisms in eukaryotes

In the Eukaryota domain, one of the three domains of life, a large number of species have separate sexes where males and females have different reproductive strategies. Why would have sex evolved and, more importantly, why has it been conserved by selection, is a very debated question. Without coming into the details, we can say here that sex is hypothesized to have been selected for in eukaryotes because it forces individuals to mate with one another, thus increasing the genetic diversity in the population. Indeed, genetic exchange by recombination between two homologous chromosomes will form a chromosome containing a different combination of alleles, which might be new in the whole population, when the genome is not inbred.

In a species with two separate sexes a way is needed to determine, for each embryo, if its gonads will develop as testes or ovaries. An extreme variety of sex determination mechanisms exists in nature. First, sex can be determined either by the environment or genetically. In an environmental determination, the conditions in which the embryo grows will influence the gonad development, thus determining the future sex of the individual. For example, in reptiles a variation in the temperature during development will determine the sex (reviewed in [1]). Second, genetic sex determination can be divided into three types: 1) By a single gene, such as the Hymenopteran *Csd* (complementary sex determiner) gene: being heterozygous for this gene will produce a female while being homozygous or single copy will make a male (reviewed in [2]). 2) By an active sex chromosome: for instance in mammals, having a Y will determine the sex as a male, because of the master sex-determining gene *Sry* [3]. 3) By the number of

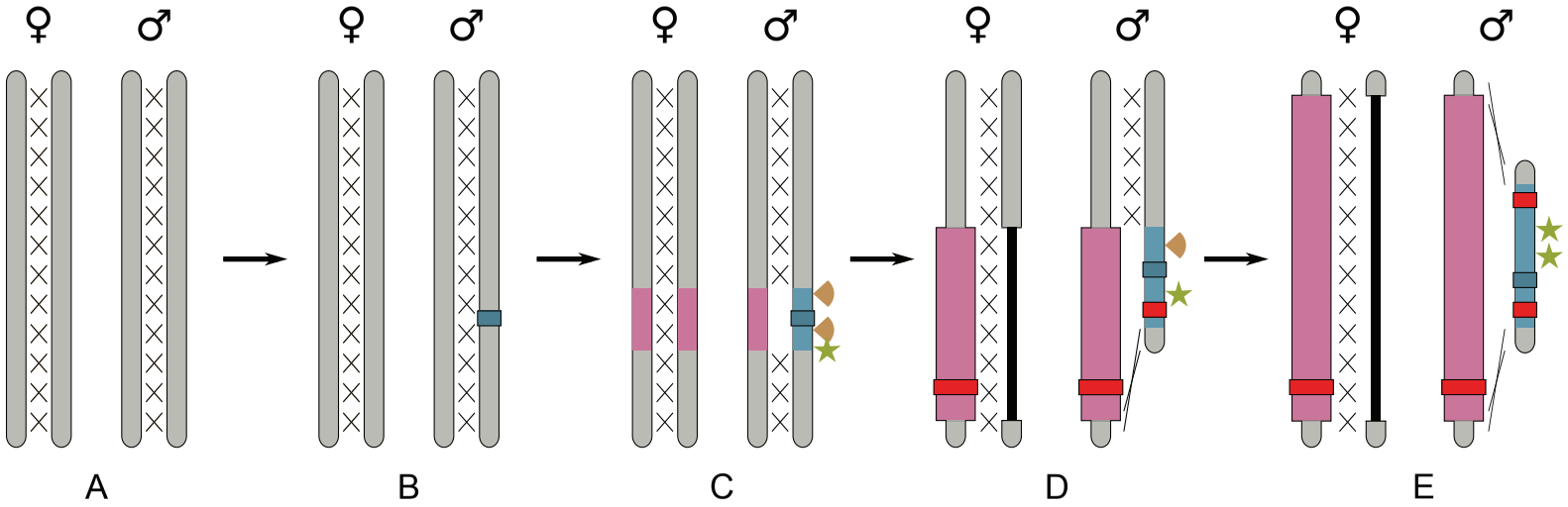
sex chromosomes compared to autosomes (X/A or Z/A): in *Drosophila melanogaster*, the presence of a unique X chromosome (with or without a Y) produces a male, whereas females have two X chromosomes (having an extra Y or not) [4]. Third, sex chromosomes are separated into three types (reviewed in [5]): 1) Sex chromosomes are called XY when the heterogametic sex (sex harboring the two different sex chromosomes) is the male, as in mammals. 2) In systems where the heterogametic sex is the female on the contrary, sex chromosomes are named ZW, as for example in birds. 3) In organisms where the sex is determined at the haploid phase, so that all diploids are heterogametic, the sex chromosomes are denoted UV. We can find this system for instance in brown algae or Bryophytes [5].

Almost all model species used in biological sciences harbor sex chromosomes: mammals, the fruit fly *Drosophila melanogaster*, the nematode *Caenorhabditis elegans*. We will thus focus from now on on this best known system of sex determination by sex chromosomes. For simplification purposes, in the rest of this introduction we will focus on XY systems unless otherwise stated.

Differenciation of X and Y chromosomes

Sex chromosomes are most of the times former autosomes that became proto-sex chromosomes when one of them (the proto-Y) acquired one or several sex-determining genes [6]. When several sex-determining genes are involved and/or when the sex-determining gene locus is close to sexually-antagonistic genes (beneficial to one sex, detrimental to the other) on the proto-Y chromosome, then a recombination arrest between the X and Y chromosomes will be selected for [7, 8, 9]. Indeed, if the different genes responsible for the sexual dimorphisms between males and females in a species are mixed, then less fit individuals will be produced. This region of recombination arrest between the proto-sex chromosomes can be small, and is indeed often small in species harboring recently evolved sex chromosomes, as is the case for instance in many dioecious plants (see e.g. [10]). The part of the proto-sex chromosomes that is still recombining is called a pseudo-autosomal region (PAR). See Figure 1A-C.

However, the current view is that most of the time the older the acquisition of the sex-determining gene is, the bigger the non-recombining region is found. This is explained by successive recombination stops following the first one, ponctually during the evolution of the sex chromosomes [11]. Again, the explanation commonly accepted for these secondary recombination arrests is a selection for a linkage between the chromosome specific to one sex and the sexually-antagonistic genes advantaging the same



General scheme of sex chromosome evolution (mammalian XY system). (A) Sex chromosomes were formaly autosomes (grey) recombining on their whole length (crosses), until **(B)** one of them acquired a sex-determining gene: here a male-determining gene on the future Y (blue rectangle). **(C)** There was selection for a first recombination arrest that formed an X-specific region (pink) and a Y-specific region (blue), the still-recombining regions being called PARs (grey). The Y started to degenerate, e.g. by pseudogenization events (green star). It also accumulated TEs (brown portion). **(D)** Further recombination stops can be selected for, if sexually-antagonistic genes appear on the sex chromosomes (red rectangle). This forms different evolutionary strata, enlarging the sex-specific regions that start degenerating on the Y, and narrowing the PARs. The Y eventually becomes smaller, by deletions and heterochromatinization. Dosage compensation evolves outside the PARs: in mammals, Ohno's hypothesis states that the X chromosome doubles its expression in both sexes, and inactivates the genes on one randomly chosen X in females (black). **(E)** Supplementary strata can be formed, enlarging the sex-specific regions and selecting for the extension of dosage compensation to these regions.

sex [8]. For example, if the Y chromosome acquires in its PAR a gene (e.g. by mutation or transposition) that is beneficial to males but detrimental to females, then selection will favor the fixation of events (e.g. inversions or mutations stopping recombination) in the population that genetically link this gene to the already non-recombining region of the Y. Without recombination, this sexually-antagonistic gene will not be transposed nor copied on the X chromosome and will thus never be in a female individual. See Figure 1D.

Whenever a region stops recombining between the X and the Y, its two allelic forms (one on the X and one on the Y) become independent from each other and start to diverge. We can then trace back the moment when they stopped sharing their evolutionary history by looking at their divergence level. One way to do this is by studying dS, the average number of neutral substitutions per site between two coding sequences. These successive recombination arrests have been shown to concern a whole region of the sex chromosomes at a time, the resulting regions of uniform divergence have thus been called "evolutionary strata" [11]. In humans, 5 strata have been described: the first two strata are ~180 million years old (myo), stratum 3 is ~100 myo, stratum 4 ~40 myo and stratum 5 ~30 myo [11, 12]. This accepted number of strata is however probably a underestimation, as recent papers tend to partition them (especially the oldest ones) [13, 14].

Effects of the absence of recombination on the Y chromosome

Because of successive recombination arrests, sex chromosomes mostly do not recombine between each other, except for the PARs (~5% of the human Y according to Ensembl assembly GRCh37) that are very often maintained. Several reasons can explain their maintenance, one of them being the proper segregation of sex chromosomes during males meiosis (reviewed in [15]). Yet, sex chromosomes enable us to study the effects of recombination on the genome, precisely because the Y chromosome is one of the most striking parts of the genome that is never recombining (the X, on the contrary, still recombines two third of the time in homogametic females). As the X and Y chromosomes were nearly identical before becoming sex chromosomes, by comparing them we can see the effects of an absence of recombination. For this we have to make the assumption that the X chromosome did not change since it became a sex chromosome, which is not completely true (see section on X changes). Still, it has been shown that the synteny of X chromosomes is well conserved both within mammals and between mammals and birds (in which another pair of sex chromosomes evolved, thus the mammalian X is

autosomal in birds) [12, 16].

Selection is less efficient on a non-recombining region comprising several loci, because of the numerous Hill-Robertson effects (reviewed in [17]). For example, when one site undergoes a highly advantageous mutation and another in the same region acquires a slightly deleterious one, if both loci are genetically linked selection will favor both mutations at a time. Thus, by "selective sweep" (one of the Hill-Robertson effects), this non-recombining region will accumulate slightly deleterious mutations [17].

Deprived of recombination for a long time, the male-specific part of the Y chromosome is highly degenerated [18, 12], due to all the different processes called Hill-Robertson effects [17]. Indeed, deleterious mutations accumulated causing pseudogenization (the human Y has lost ~97% of its gene content [18]), reduced expression of the remaining genes and less frequent optimal codons usage (studied in *Drosophila*: [19, 20]). Other types of events than mutations are also less efficiently removed from the Y chromosomes: it accumulates repeated and transposable elements (TEs) [21, 22], and some of the inversions that led to the formation of secondary strata may have not been selected for but simply not purged from the population. See Figure 1C-E.

Gene conversion on the Y chromosome

Despite all this degeneration, which led scientists to suppose for long that the Y chromosome is a desolate chromosome, this sex chromosome is generally not lost from the genome and it does retain some of its gene content during evolution. This constituted a puzzle, until a certain type of recombination was discovered to occur on the Y chromosome: gene conversion [23, 18]. Gene conversion is a type of recombination happening during both crossing over (CO) and non-CO events, where a non-reciprocal transfer of genetic information occurs: a "donor" copies its sequence on the homologous "receptor" after this latter underwent a double-strand break (see e.g. [24]).

Intra-Y chromosome (Y-Y) gene conversion happens in the so-called ampliconic regions [25, 18]. In these regions, multicopy genes are present, that share homology with an X-linked gene but have a testis-specific expression [18]. Non-CO events can occur between these different copies of a same Y-linked gene. After a double-strand break occurs in one copy, due to the presence of other copies of this gene in the vicinity, the recombination machinery will repair the break by copying one of these copies on the damaged locus. As Y-Y gene conversion between copies is a form of recombination, it slows down Muller's ratchet and thus improves the efficacy of selection [17]. This mechanism is thus probably responsible for the maintenance of essential Y-linked genes.

However the human and chimpanzee Ys do not share all of their amplicons [26], thus these genes may have been conserved by other means earlier during evolution.

Gene conversion has also been shown to occur between the X and Y gametologs (homologs that were allelic when the sex chromosomes where still autosomes) [23, 27, 28]. Despite having been less studied, this X-Y gene conversion has been also hypothesized to help maintaining Y-linked genes [23, 29, 28]. This time, the template used to repair a double strand break on a Y-linked gene will be its X gametolog, on which selection has always been efficient.

Evolutionary changes on the X chromosome

During sex chromosome evolution, the X and Y chromosomes diverge, mainly because the Y chromosome degenerates. But the X chromosome is not an impassive witness of its partner degeneration: despite its strong conservation in terms of gene content, the X chromosome acquired specific features [30, 31, 32]. For instance, the mammalian X chromosome got enriched in genes involved in both brain developpment and sexual dimorphism [12]. Yet the most impressive X specialization is probably dosage compensation. Dosage compensation is hypothesized to be the response of the X chromosome to the differentiation of its partner (e.g. by gene loss). When sex chromosomes were still autosomes, they had two allelic copies of a same gene, as is the case for all autosomes in a diploid genome. But for each Y-linked gene that has been lost, or that differentiated into something different (either in terms of proteic function or of tissues expression), the corresponding X-linked gene became single copy. This led to a problem of gene dosage in XY males.

X dosage compensation mechanisms have been found to be very different between the studied species. In *Drosophila melanogaster*, epigenetic signals double the expression of most of the X-linked genes specifically in males ([33] and references therein). This mechanism allows to resolve the dose problem directly in the only sex that needs it: the males. In mammals, no evidence for a doubling of expression has been provided for a long time, meaning that contrarily to *Drosophila* there is no sex-specific global expression change. However, the inactivation of one of the two X chromosomes in females was discovered independently by Susumu Ohno and Mary Lyon [34, 35]. This X-chromosome inactivation (XCI) is a global mechanism that affects gene expression, thus it has been interpreted as dosage compensation. However, the evolution of XCI was difficult to explain: rather than solving the expression problem in males it seems to extend it to females. Ohno proposed that dosage compensation in mammals evolved as a

two-step mechanism with 1) a two-fold expression increase of the X chromosome in both sexes, which solves the gene dose imbalance problem in males, and 2) inactivation of one of the two X chromosomes by XCI to restore optimal dosage in females [6]. Lacking genomic technologies allowing a proper estimation of gene expression, this hypothesis remained for long untested though widely accepted in the scientific community. See Figure 1D-E.

Unresolved questions

The major hypotheses on sex chromosome evolution described above have been developped since the 60s, by using genetic data [6, 36, 37]. With the help of genomic data, these hypotheses were confirmed and extended in the 2000s [11, 18, 12]. One of the major challenges of the field is to test these hypotheses further, by using the current incredibly fast production of genomic data.

During my thesis, I focused on different questions on sex chromosome evolution. First, I studied how the Y chromosome is rescued by the X, i.e. what mechanisms does the X establish as a response to Y degeneration. I made two different studies:
(1) As said above, the first part of Ohno's hypothesis, i.e. the putative doubling of expression of all X-linked genes in both sexes, was never tested until the 2000s. In 2006, two studies used microarray data and found evidence for this two-fold expression increase [38, 39], thus validating Ohno's hypothesis. But in 2010, using RNAseq data which are more accurate to study differences in expression between genes of a same cell, Xiong and collaborators claimed that Ohno's hypothesis was wrong: they did not find any expression increase on the X compared to the autosomes [40]. An avalanche of papers followed, from either proponents or opponents of Ohno's hypothesis. My results show that Ohno was not completely right, as the arguments used by Ohno's proponents to prove there is a global X hyperexpression exhibit some caveats. Nor was he completely wrong, as some X-linked genes indeed display the hypothesized two-fold expression increase. As the situation was very confused with this controversy about Ohno's hypothesis, and also because my results and others called for a revision of the link between global XCI and local (gene-by-gene) X hyperexpression, I participated in the writing of a review on the evolution of dosage compensation in mammals.
(2) During the last decade, some authors hypothesized that the X might help maintaining Y-linked genes by gene conversion [23, 29, 28], as explained above. However, X-Y gene conversion has only been studied in very few loci and by comparing few species. Using data produced by my collaborators in Toulouse, I studied X-Y gene conversion in

five gametologs spanning three different human evolutionary strata, in a set of species distributed in the Simian phylogeny. My results suggest that there is no strong evidence for X-Y gene conversion acting against Y degeneration, and that its maintenance in primates might be neutral.

Second, as the models of sex chromosome evolution exposed above are based on the study of very few model species, it is necessary to test them in non-model species. Moreover, of the three types of sex chromosomes introduced above, the haploid UV system is the least studied. In this system, sex is determined during the haploid phase: males are V, females are U (sexes being determined by the size of the gametes produced), and diploids are all UV. There is thus no homogametic sex: both the V and U chromosomes are alone when the sex is expressed. This leads to several expectations on the evolution of this type of sex chromosome, different from the ones for XY and ZW systems [41, 5]. As a member of a consortium, I studied the brown alga *Ectocarpus siliculosus* which possesses UV sex chromosomes and, worth noticing, is from a eukaryotic supergroup distantly related to that of the well studied sex chromosomes of animals and plants (>1 billion years of divergence). This species also displays little differentiated sexes (in between isogamy and oogamy), allowing us to test if the general rules of sex chromosome evolution found in animals still hold for species having little sexual dimorphism. Our results show that the general scheme of sex chromosome evolution applies in this species as well: both sex-specific regions are very probably former autosomes that stopped recombining, accumulated TEs, probably underwent gene loss (as gene density is significantly lower than on the autosomes), and conserved some gametologs. We were able to estimate the age of this sex chromosome system between 100 and 200 million years old. Yet, these U and V chromosomes are quite homomorphic as the sex-specific regions represent one fifth of their length. This questions the common view that old sex chromosomes (either XY, ZW or UV) have to be heteromorphic (together with studies on ratite birds [42]).

Bibliography

[1] Merchant-Larios H, Díaz-Hernández V (2013) Environmental sex determination mechanisms in reptiles. Sex Dev 7: 95–103.

[2] Heimpel GE, de Boer JG (2008) Sex determination in the hymenoptera. Annu Rev Entomol 53: 209–230.

[3] Sinclair AH, Berta P, Palmer MS, Hawkins JR, Griffiths BL, et al. (1990) A gene from the human sex-determining region encodes a protein with homology to a conserved dna-binding motif. Nature 346: 240–244.

[4] Cline TW, Meyer BJ (1996) Vive la différence: males vs females in flies vs worms. Annu Rev Genet 30: 637–702.

[5] Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, et al. (2011) Are all sex chromosomes created equal? Trends Genet 27: 350–357.

[6] Ohno S, et al. (1967) Sex chromosomes and sex-linked genes.(Monographs on endocrinology, Vol. 1.). Berlin, Heidelberg, New York: Springer Verlag. URL `http://www.cabdirect.org/abstracts/19680100985.html`.

[7] Rice WR (1987) The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution 41: 911–914.

[8] Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb) 95: 118–128.

[9] Bachtrog D (2013) Y-chromosome evolution: emerging insights into processes of y-chromosome degeneration. Nat Rev Genet 14: 113–124.

[10] Ming R, Bendahmane A, Renner SS (2011) Sex chromosomes in land plants. Annu Rev Plant Biol 62: 485–514.

[11] Lahn BT, Page DC (1999) Four evolutionary strata on the human x chromosome. Science 286: 964–967.

[12] Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al. (2005) The dna sequence of the human x chromosome. Nature 434: 325–337.

[13] Wilson MA, Makova KD (2009) Evolution and survival on eutherian sex chromosomes. PLoS Genet 5: e1000568.

[14] Pandey RS, Wilson Sayres MA, Azad RK (2013) Detecting evolutionary strata on the human x chromosome in the absence of gametologous y-linked sequences. Genome Biol Evol 5: 1863–1871.

[15] Otto SP, Pannell JR, Peichel CL, Ashman TL, Charlesworth D, et al. (2011) About par: the distinct evolutionary dynamics of the pseudoautosomal region. Trends Genet 27: 358–367.

[16] Ezaz T, Stiglec R, Veyrunes F, Marshall Graves JA (2006) Relationships between vertebrate zw and xy sex chromosome systems. Curr Biol 16: R736–R743.

[17] Charlesworth B, Charlesworth D (2000) The degeneration of y chromosomes. Philos Trans R Soc Lond B Biol Sci 355: 1563–1572.

[18] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.

[19] Steinemann M, Steinemann S (1998) Enigma of y chromosome degeneration: neo-y and neo-x chromosomes of drosophila miranda a model for sex chromosome evolution. Genetica 102-103: 409–420.

[20] Bartolomé C, Charlesworth B (2006) Evolution of amino-acid sequences and codon usage on the drosophila miranda neo-sex chromosomes. Genetics 174: 2033–2044.

[21] Charlesworth B, Sniegowski P, Stephan W (1994) The evolutionary dynamics of repetitive dna in eukaryotes. Nature 371: 215–220.

[22] Kvikstad EM, Makova KD (2010) The (r)evolution of sine versus line distributions in primate genomes: sex chromosomes are important. Genome Res 20: 600–613.

[23] Pecon Slattery J, Sanner-Wachter L, O'Brien SJ (2000) Novel gene conversion between x-y homologues located in the nonrecombining region of the y chromosome in felidae (mammalia). Proc Natl Acad Sci U S A 97: 5307–5312.

[24] de Massy B (2003) Distribution of meiotic recombination sites. Trends Genet 19: 514–522.

[25] Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, et al. (2003) Abundant gene conversion between arms of palindromes in human and ape y chromosomes. Nature 423: 873–876.

[26] Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, et al. (2010) Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. Nature 463: 536–539.

[27] Marais G, Galtier N (2003) Sex chromosomes: how x-y recombination stops. Curr Biol 13: R641–R643.

[28] Trombetta B, Cruciani F, Underhill PA, Sellitto D, Scozzari R (2010) Footprints of x-to-y gene conversion in recent human evolution. Mol Biol Evol 27: 714–725.

[29] Rosser ZH, Balaresque P, Jobling MA (2009) Gene conversion between the x chromosome and the male-specific region of the y chromosome at a translocation hotspot. Am J Hum Genet 85: 130–134.

[30] Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, et al. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 6: e80.

[31] Bellott DW, Skaletsky H, Pyntikova T, Mardis ER, Graves T, et al. (2010) Convergent evolution of chicken z and human x chromosomes by expansion and gene acquisition. Nature 466: 612–616.

[32] Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M (2010) Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the x chromosome. PLoS Biol 8.

[33] Disteche CM (2012) Dosage compensation of the sex chromosomes. Annu Rev Genet 46: 537–560.

[34] Ohno S S, Kaplan WD, Kinosita R (1959) Formation of the sex chromatin by a single x-chromosome in liver cells of rattus norvegicus. Exp Cell Res 18: 415–418.

[35] Lyon MF (1961) Gene action in the x-chromosome of the mouse (mus musculus l.). Nature 190: 372–373.

[36] Nei M (1969) Linkage modifications and sex difference in recombination. Genetics 63: 681–699.

[37] Charlesworth B, Charlesworth D (1978) A model for the evolution of dioecy and gynodioecy. American naturalist 112: 975–997.

[38] Nguyen DK, Disteche CM (2006) Dosage compensation of the active x chromosome in mammals. Nat Genet 38: 47–53.

[39] Gupta V, Parisi M, Sturgill D, Nuttall R, Doctolero M, et al. (2006) Global analysis of x-chromosome dosage compensation. J Biol 5: 3.

[40] Xiong Y, Chen X, Chen Z, Wang X, Shi S, et al. (2010) Rna sequencing shows no dosage compensation of the active x-chromosome. Nat Genet 42: 1043–1047.

[41] Bull JJ, et al. (1983) Evolution of sex determining mechanisms. The Benjamin/Cummings Publishing Company, Inc.

[42] Ogawa A, Murata K, Mizuno S (1998) The location of z- and w-linked marker genes and sequence on the homomorphic sex chromosomes of the ostrich and the emu. Proc Natl Acad Sci U S A 95: 4415–4418.

1

Rescuing from the X: the evolution of dosage compensation in mammals

Before the beginning of my PhD, Gabriel Marais developped a project on studying dosage compensation in humans focusing on dosage-sensitive genes. A part of the project was about identifying the genes underlying X-aneuploidy syndromes. The XCI-escapees are the only X-linked genes affected by changes in dosage due to X-aneuploidies and presumably the dosage-sensitive ones among those are the best candidates. Gabriel contacted Aoife McLysaght who had recently published a paper on an approach for identifying dosage-senstive genes (dosage-balanced ohnologs, DBOs) in humans and showed that DBOs on the human chromosome 21 were most probably the genes underlying the Down syndrome (Makino & McLysaght *PNAS* 2010).

I started the analysis with receiving data on DBOs and protein-complexes in humans from Takashi Makino and Aoife McLysaght. In December 2010, the publication of a paper in *Nature Genetics* by Xiong *et al.* challenging Ohno's hypothesis made us explore the possiblity that only dosage-sensitive genes were hyperexpressed on the X chromosome. Computing the X:AA expression ratio of X-linked DBOs resulted in values around 0.5, meaning they are not globally twice more expressed on the X than on autosomes. Thus, rather than comparing X-linked and autosomal DBOs as a whole, we decided to study another type of dosage-sensitive genes: members of proteic complexes. This allowed us to compare X:AA ratios inside clusters of genes whose expressions are

supposed to be constrained together.

I presented a poster of the obtained results in three different international conferences:

- "Genetics, Epigenetics and Evolution of Sex chromosomes", French Society of Genetics. Paris, June 2011.
- "Theoretical and empirical advances in evolutionary genomics", Jacques Monod Conference. Roscoff, April 2012.
- SMBE Annual meeting. Dublin, June 2012.

This paper was sent to *PNAS* for review on the 13^{th} of October 2011, accepted on the 3^{rd} of February 2012, and published "from the cover" on the 3^{rd} of April 2012.
A commentary from Alison Wright and Judith Mank was published on the same *PNAS* issue.

Mammalian X chromosome inactivation evolved as a dosage-compensation mechanism for dosage-sensitive genes on the X chromosome

Eugénie Pessia[a], Takashi Makino[b,c], Marc Bailly-Bechet[a], Aoife McLysaght[c], and Gabriel A. B. Marais[a,d,1]

[a]Laboratoire de Biométrie et Biologie évolutive, Université Lyon 1, Centre National de la Recherche Scientifique, Villeurbanne F-69622 cedex, France; [b]Department of Ecology and Evolutionary Biology, Graduate School of Life Sciences, Tohoku University, Aoba-ku, Sendai 980-8578, Japan; [c]Smurfit Institute of Genetics, University of Dublin, Trinity College, Dublin 2, Ireland; and [d]Instituto Gulbenkian de Ciência, P-2780-156 Oeiras, Portugal

Edited* by Michael Freeling, University of California, Berkeley, CA, and approved February 3, 2012 (received for review October 13, 2011)

How and why female somatic X-chromosome inactivation (XCI) evolved in mammals remains poorly understood. It has been proposed that XCI is a dosage-compensation mechanism that evolved to equalize expression levels of X-linked genes in females (2X) and males (1X), with a prior twofold increase in expression of X-linked genes in both sexes ("Ohno's hypothesis"). Whereas the parity of X chromosome expression between the sexes has been clearly demonstrated, tests for the doubling of expression levels globally along the X chromosome have returned contradictory results. However, changes in gene dosage during sex-chromosome evolution are not expected to impact on all genes equally, and should have greater consequences for dosage-sensitive genes. We show that, for genes encoding components of large protein complexes (≥ 7 members)—a class of genes that is expected to be dosage-sensitive—expression of X-linked genes is similar to that of autosomal genes within the complex. These data support Ohno's hypothesis that XCI acts as a dosage-compensation mechanism, and allow us to refine Ohno's model of XCI evolution. We also explore the contribution of dosage-sensitive genes to X aneuploidy phenotypes in humans, such as Turner (X0) and Klinefelter (XXY) syndromes. X aneuploidy in humans is common and is known to have mild effects because most of the supernumerary X genes are inactivated and not affected by aneuploidy. Only genes escaping XCI experience dosage changes in X-aneuploidy patients. We combined data on dosage sensitivity and XCI to compute a list of candidate genes for X-aneuploidy syndromes.

Y degeneration | sex-linked gene expression | balance hypothesis

The sex chromosomes of therian mammals (placentals and marsupials) originated from a pair of autosomes about 150 million years ago (1–5). The X and Y chromosomes gradually diverged after several events of recombination suppression, probably inversions on the Y chromosome (3, 6, 7). With the exception of two very small pseudoautosomal regions (PARs), the Y chromosome never recombines. Because of its nonrecombining nature, the Y chromosome has degenerated and lost most of its genes (reviewed in ref. 8). In contrast, during therian evolution the recombining X chromosome has retained many ancestral genes (6), gained new genes, and evolved new expression patterns for some genes (4, 9, 10).

Early in the differentiation of the sex chromosomes, most ancestral genes were present on both X and Y and the imbalance of gene products between males and females would have been small. As the attrition of Y chromosome genes progressed, an increasing number of loci were uniquely present on the X chromosome, implying a twofold reduction of their expression in males (XY) compared with females (XX). X chromosome inactivation (XCI) in females makes expression of X-linked genes similar in males and females (11). However, instead of solving the problem of dosage imbalance between autosomal and X genes, XCI seemed to expand it to females. Ohno proposed that the global expression of the X chromosome must have doubled in both sexes during evolution, solving the X:autosome imbalance in males, and suggested that XCI had evolved subsequently to reduce the output of X-linked genes back to the ancestral levels in females (1). Both steps are required to call XCI a dosage-compensation mechanism.

Consistent with Ohno's hypothesis, microarray data suggested that the mammalian X chromosome global expression level was similar to that of autosomes (12–14). However, analysis of RNA-seq data, which yield much more precise expression-level estimates than microarray data, indicated that X chromosome global expression level in humans and mice was half that of autosomes in both sexes (15). This analysis suggested that the first step in Ohno's scenario was missing and raised doubt about XCI as a dosage-compensation mechanism (16). It was recently shown that this conclusion was strongly affected by the inclusion of testis-specific X genes in the analysis (17) (see also refs. 18–22 about the controversy regarding dosage compensation in mammals). However, even though excluding these genes with no expression in somatic tissues brings X chromosome global expression closer to that of autosomes, it is still significantly lower, suggesting that the true nature of X dosage compensation may differ from these "all or nothing" scenarios.

In zebra finch, chicken, and crow, partial dosage compensation has been observed on the avian Z chromosome (23–26), with only some Z-linked genes being dosage-compensated (27, 28). Partial Z chromosome dosage compensation has also been observed in silkworm (29, 30) and in the parasite *Schistosoma mansoni* (31). Studies on the platypus indicate incomplete X chromosome dosage compensation in monotremes (32, 33). Data from sticklebacks show that the X is more strongly expressed in females than in males (34), consistent with a lack of global sex-chromosome dosage compensation in this fish. All this work suggests that global dosage compensation might not be a general feature of sex chromosomes (35).

Changes in gene dosage during sex-chromosome evolution are only expected to affect dosage-sensitive genes (35), which could explain why partial dosage compensation has been observed when analyzing all X/Z genes combined together. In this study, we focused on dosage-sensitive genes in the human genome and tested for dosage compensation of these genes only. Early experiments comparing polyploids and aneuploids in plants have shown that imbalanced expression of dosage-sensitive genes can strongly impact the phenotype (36). It was later shown that in yeast, most dosage-sensitive genes are involved in protein complexes (37). Using experimental data from strains heterozygous for single-gene knockouts, Papp et al. (37) could indeed show that a strong decline in fitness is only observed for genes encoding proteins

Author contributions: G.A.B.M. designed research; E.P., T.M., and M.B.-B. performed research; T.M. and A.M. contributed new reagents/analytic tools; E.P., M.B.-B., and G.A.B.M. analyzed data; and E.P., A.M., and G.A.B.M. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

See Commentary on page 5144.

[1]To whom correspondence should be addressed. E-mail: gabriel.marais@univ-lyon1.fr.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1116763109/-/DCSupplemental.

involved in complexes (hereafter named protein-complex genes), such as the ribosome. They also found that proteins from the same complexes tend to be coexpressed at very similar levels and tend to have the same number of copies. All these lines of evidence, and others, suggest that there are strong constraints on the stoichiometry of the members of a complex and that an imbalance in such stoichiometry can be deleterious (37).

This "balance hypothesis" has become very popular and has been repeatedly used to explain patterns of duplicate gene evolution in yeast, *Arabidopsis*, rice, composites, and *Paramecium* (reviewed in refs. 38 and 39). In these organisms, most duplicate genes maintained after whole-genome duplication (WGD) events over large evolutionary periods are involved in protein complexes (40–45). In contrast, protein-complex genes are underrepresented in duplicates from segmental duplications (46, 47). These patterns of gene duplicability are fully predicted by the balance hypothesis because WGD events will not affect the stoichiometry of the components of a complex (but subsequent loss of a single component will be counter-selected), whereas segmental duplications will disrupt the stoichiometry.

It has been suggested that in multicellular organisms, selection for balanced dosage may be weaker than in unicellular organisms because selection is reduced in such organisms with small effective population size (48, 49). Additionally, genes involved in regulatory networks (such as transcription factors) are also expected to be dosage-sensitive, and in multicellulars these genes are probably numerous (39). However, in multicellulars, many dosage-sensitive genes are likely to be protein-complex genes. This idea was explored in humans and dosage-sensitive genes were identified as genes maintained after WGD events and resistant to segmental duplications and copy-number variations, and called dosage-balanced ohnologs (DBOs) (50). Protein-complex genes were found overrepresented among these DBO genes. Strikingly, 75% of the Down syndrome (trisomy 21) candidate genes are DBOs and a highly significant excess of DBOs was found in the Down syndrome critical region, which is known as a major determinant of the features of this syndrome. This finding is consistent with the observation that many haploinsufficient genetic diseases in humans are caused by protein-complex genes (51).

Here we focused on protein-complex genes in humans to test for the evolution of dosage compensation in dosage-sensitive X-linked genes. We used a list of protein-complex genes inferred from experimental data and expression-level estimates from RNA-seq data in humans. Based on these results we built a list of genes of interest for X aneuploidy syndromes, our rationale being that dosage-sensitive genes that escape X inactivation could be a cause of the phenotypes observed in these syndromes.

Results and Discussion

Expression Analysis of Dosage-Sensitive X Genes and Evidence for Dosage Compensation in Humans. Assuming global autosomal expression level has not changed since the X and Y chromosomes originated, and if XCI has evolved to make autosomal and X expression equal (dosage compensation), as in Ohno's scenario, then the X/A mean expression ratio should be 1 in both males and females. A value of 0.5 in males is expected if X expression has remained constant along XY chromosome evolution, because males only have one copy of the X chromosome. In this case, a value of 0.5 is also expected in females because one of the two X chromosomes is inactivated and not transcribed. An X/A expression ratio of 0.5 would mean that XCI does not act as a dosage-compensation mechanism and its role is equivocal. Our independent analysis of data from 12 male and female tissues from ref. 15 found, as did the authors of the original analysis of that dataset, that X chromosome global expression is about half the expression of autosomes in both male and female (Fig. 1*A*). We checked whether differences in dataset/raw read processing could explain differences in conclusions found in Xiong et al. (15) and Deng et al. (17), but found expression-level estimates from both studies to be strongly correlated (*Materials and Methods*). Using data from Xiong et al. (15), we found that the

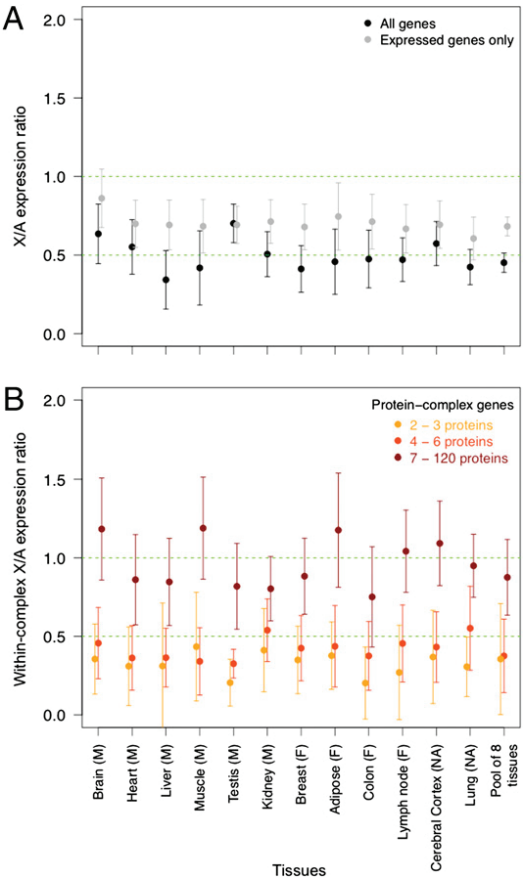


Fig. 1. X/A expression ratio. (*A*) In this analysis, 734 X genes and 19,066 autosomal genes are included (*Materials and Methods*). Expression of X genes is normalized by the median of autosomal gene expression. The median of X/A ratios and associated 95% confidence interval are shown for each tissue. Results for both all genes (black, as in ref. 15) and excluding nonexpressed genes (gray, as in ref. 17) are shown. (*B*) Here only genes involved in protein complexes are included. For each complex, we computed the median of X gene expression over that of autosomal gene expression. We prepared three groups with similar sample size with increasing protein-complex size in number of proteins: small (2–3 proteins, yellow), medium (4–6 proteins, orange), large (7–120 proteins, brown). For each tissue and complex size category, the median of within-complex X/A ratios and associated 95% confidence interval are shown. In both panels we show the results for a pool of eight tissues (see text). The two green dashed lines indicate expectations with dosage compensation (X/A = 1) and without dosage compensation (X/A = 0.5).

X/A expression ratio significantly increases when the nonexpressed genes are removed, as in Deng et al. (17): it is now close to 0.7 for most of the tissues (Fig. 1*A*). Such an increase is explained by a higher fraction of nonexpressed genes on the X chromosome than on the autosomes; indeed, the X chromosome includes many testis-specific genes not expressed in somatic tissues in both humans and mice (17, 19, 22). Taken together, these data suggest that our work is not affected by differences in datasets or procedure for data-analysis, as we are able to make the same observations as in refs. 15 and 17 using different filters.

Deng et al. (17) inferred that there is a global up-regulation of X gene expression in humans (see also ref. 19). For most tissues, however, the X/A expression ratio is close to 0.7 and is not 1, the expected value for global X up-regulation (Fig. 1*A*). Deng et al. (17) suggested that this is because RNA-seq data are noisy: genes with RNA-seq–estimated low expression levels can actually be nonexpressed genes (see also ref. 19). As these genes are comparatively more numerous on the X than on the autosomes, an X/A expression ratio smaller than 1 is expected from noisy RNA-

seq data. Using brain as an example, Deng et al. (17) argued that when the expression-level distributions from X and autosomes are compared, they seem to be similar, which supports global X up-regulation (see figure 1*A* in ref. 17). However, many tissues do show nonoverlapping X and autosomal distributions (see supplementary figure 1 in ref. 17). Instead, we interpret these data as suggestive that the X chromosome includes a mixture of up-regulated and nonup-regulated genes, the combined analysis of which returns a "mixed" X/A expression ratio between 0.5 (no dosage compensation) and 1 (full dosage compensation).

To refine the analysis we examined genes involved in protein complexes from the Human Protein Reference Database (HPRD) list (*Materials and Methods*) because these genes are likely to be dosage-sensitive and should be the main target for dosage compensation. This list includes 207 human protein complexes with proteins from both X and autosomal genes. We computed the X/A expression ratio within each complex and we obtained the median of this ratio among all complexes present in a given tissue, as well as pooled for all tissues excluding "reproductive" organs (testis, breast) and brain/nervous system tissues, because X-linked genes are known to be overrepresented and overexpressed in these tissues (6, 15) (Fig. 1). A preliminary analysis showed that complex size is a strong determinant of the X/A expression ratio within protein complexes, as revealed by a multiple regression analysis, including complex size, whole-complex gene expression, and percentage of X proteins in a complex [effects on X/A from strongest to weakest: complex size, P value = 0.0007; percentage of X proteins, P value = 0.03; global complex expression, nonsignificant]. In Fig. 1*B*, we therefore showed the results for three protein-complex size categories, each containing the same number of complexes. This analysis clearly shows that the X/A expression ratio increases with protein-complex size. For large protein complexes (≥7 proteins), the X/A expression ratio is significantly higher than 0.5 for 11 of 12 tissues, as well as in the pooled expression data. For most tissues (10 + pooled expression data), the X/A expression ratio is not significantly different from 1, the value expected in case of dosage compensation. Our observation suggests that dosage-sensitivity is stronger for genes involved in large protein complexes than for genes involved in small protein complexes, which makes dosage compensation required more often for the former. Complex size has also been found to have some influence on dosage sensitivity in yeast, because the fitness effect of dosage imbalance in heterozygous knockout mutants is correlated to protein-complex size (37). Several explanations could account for this finding. First, if imbalance leads to incomplete (and nonfunctional) complexes that are destroyed by the cell, the bigger the complex, the bigger the metabolic cost for the cells. Second, subunits forming a bridge between parts of the complex can inhibit complex assembly if present in excess. This problem should increase with the number of subunits in a complex.

Importantly, our results are unaffected by inclusion or exclusion of nonexpressed genes (the patterns shown in Fig. 1*B* remain exactly the same after removal of nonexpressed genes) (Fig. S1). Deng et al. (17) showed that the X/A expression ratio increased from 0.5 to 1, also increasing the minimum expression-level threshold required for a gene to be included in the analysis (see figure 1*D* in ref. 17; see also ref. 19). These authors suggested this finding was because of noise in the RNA-seq data affecting lowly expressed genes (see above) and concluded that the X chromosome was probably up-regulated as a whole. Using a similar test, we found that when the X/A expression ratio reaches 1, only a small fraction of X genes (159; i.e., 21% of the initial set of X genes) are still being analyzed, which weakens the idea of global up-regulation on the X chromosome (Fig. S2). We also found that the fraction of protein-complex genes increases when applying different thresholds for gene expression and moving the X/A expression ratio from 0.5 to 1. Importantly, this pattern is stronger for large complexes than for other complexes (P value = 0.034, Fisher's exact test with two categories for expression level using data from ref. 15: ≤ 0.05 and > 0.05), which suggests highly expressed genes include more dosage-sensitive genes, a trend that has been noted before (52). When excluding lowly expressed genes from the dataset, we may be getting rid of noisy data but we also seem to be enriching the dataset in dosage-sensitive genes, which could help getting an X/A expression ratio of 1 in agreement with up-regulation of the X chromosome affecting mostly dosage-sensitive X genes.

A balanced X/A expression ratio for a given complex could result from a twofold increase of X gene expression (as in Ohno's scenario) to match autosomal gene expression, or a twofold reduction of autosomal gene expression to match X gene expression. To distinguish these possibilities, we split our dataset in two categories: large complexes (≥ 7 proteins; L) and other complexes (< 7 proteins; O) and computed the ratio of the expression in large complexes and in other complexes separately for X genes (X_L/X_O ratio) and for autosomal genes (A_L/A_O ratio). We found that the A_L/A_O ratio is close to 1 for all tissues (Fig. 2 and Fig. S1). This result indicates that expression of autosomal genes does not differ significantly between large and other complexes, so the dosage changes we observed (Fig. 1*B*) are not a general feature of high expression levels of large complexes; rather, this feature is restricted to the X chromosome. Only about 100 X genes are included in the computation of X_L/X_O ratio (Fig. 2 legend) and, as expected for such a small dataset, the error bars are large. Nevertheless, the X_L/X_O ratio is significantly higher than 1 for 9 of 12 tissues (P value = 4.9×10^{-4}, Wilcoxon paired test) and in seven cases, this ratio is close to two, suggestive of a doubling of expression levels in agreement with dosage compensation. This doubling could be explained by an enrichment of RNA polymerase II in 5′ of the X genes compared with autosomal genes (17, 22) mediated by active histone marks (22).

We thus find evidence that the scenario put forward by Ohno of a twofold increase of X gene expression in both sexes plus inactivation of one X in female as a way of compensating for Y gene loss is valid for dosage-sensitive genes in humans. Our analysis focused on protein-complex genes, which are considered

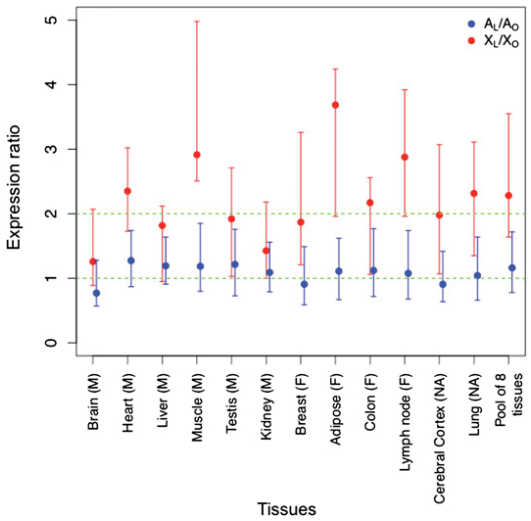


Fig. 2. X expression and autosomal expression in large protein complexes versus others. For each tissue, we computed the ratio of the median of X gene expression of large complexes (≥ 7 proteins, n = 59) and the median of X gene expression of other protein complexes (< 7 proteins, n = 52), which we called the X_L/X_O ratio (red). Both categories have been defined from results presented in Fig. 1*B*. The ratio of the median of autosomal gene expression of large complexes (n = 696) and the median of autosomal gene expression of other protein complexes (n = 151)—the A_L/A_O ratio (blue)—was computed similarly. Error bars have been obtained by bootstrapping protein complexes and computing both ratios and represent 95% bootstrap confidence interval. We pooled the data for eight tissues (see text) and computed the median and confidence interval the same way. The two green dashed lines indicate expectations with a twofold increase of expression (ratio of 2) and without any change in expression (ratio of 1).

the main source of dosage-sensitive genes in yeast (37). In multicellulars, genes involved in regulatory networks may be another major source of dosage-sensitive genes (39). We know that the dosage of some X genes escaping XCI can modulate autosomal gene expression, although the effect is small (53). This finding suggests that many dosage-sensitive regulatory X genes may be compensated, and it would be interesting to test for dosage compensation in these genes.

Dosage-Sensitive XCI Escapees as Candidate Genes for X Aneuploidy Syndromes. Most autosomal aneuploidies are nonviable, with the notable exception of chromosome 21. Interestingly, chromosome 21 is the human chromosome with the lowest number of dosage-sensitive genes, which suggests dosage-sensitive genes are key elements of aneuploidy phenotypes (50). The X aneuploidy in humans is known to have only mild effects, which at first sight may be surprising given the size and the number of genes of the X chromosome; aneuploidies of autosomes of equivalent size are all lethal. Sex-chromosome aneuploidies have a very high prevalence in humans, with Klinefleter (XXY) being the most common aneuploidy in men (1/500–600), Triple-X (XXX) being the most common in females (1/1,000), and Turner (X0) being quite common in females (1/2,000–2,500). This finding is explained by X-inactivation of all of the supernumerary X chromosomes, which means that in case of loss of one X or the presence of extra X chromosomes, only one X chromosome will be active, as in XX females (54, 55). Some genes, however, escape XCI and it has been proposed long ago that these genes could underlie Turner, Klinefelter, and other X aneuploidy syndromes (56–60). About 100 XCI escapees are currently known in humans from experiments on about 600 X genes (these are two-thirds of the X genes), which means that maybe about 150 X genes could escape XCI in total (58). This small number of genes could explain why X aneuploidies have an even milder effect than chromosome 21 trisomy (there are 223 genes on chromosome 21). Interestingly, in mice only 3% of X genes escape XCI, compared with 15% in humans, and X monosomy in mice has smaller phenotypic effects than in human, which is consistent with XCI escapees underlying X aneuploidy syndromes (61).

Dosage is clearly central in Klinefelter syndrome, as the neurodevelopmental and psychological features of patients become more severe as the number of supernumerary X chromosomes increases, for example in XXXY and XXXXY males (60). Very few candidate genes are known for any X aneuploidy syndrome. One well-established candidate gene is *SHOX*, a gene from PAR1 that is involved in small stature in Turner syndrome (62, 63). *SHOX* is haploinsufficient in Turner patients. In Klinefelter patients, *SHOX* escapes XCI and is overdosed and the prototypic Klinefelter patient is tall, which is consistent with *SHOX* being a Klinefelter gene (59, 60). The case of *SHOX* suggests that the same genes could underlie Klinefelter, Turner, and other X aneuploidy syndromes, which would make sense as these syndromes often relate to the same traits (e.g., stature, cognition). Another somewhat equivocal candidate for Turner syndrome is *RPS4* (57, 63). *RPS4* escapes XCI, has a functional Y homolog, and is located in Xq. This gene encodes a ribosomal protein and clearly falls in our dosage-sensitive gene category. Interestingly, it has been shown that 46,X,i(Xq) karyotype (i.e., isochromosome Xq) cannot be differentiated phenotypically from 45,X Turner syndrome patients (64). This finding was initially considered evidence that Turner syndrome genes are on Xp because Xp is missing in 46,X,i(Xq) patients. However, the 46,X,i(Xq) patients carry three copies of the *RPS4*X gene and the above-mentioned results are also consistent with overdosage of *RPS4* being as deleterious as half-dosage, which fits well with the dosage-balance hypothesis. The case of *RPS4* shows that dosage-sensitive genes may have similar phenotypic effects in Turner, Klinefelter, and other X aneuploidy syndromes, but other genes, such as *SHOX*, may have opposed phenotypic effects depending on gene dosage.

Our results on X chromosome protein-complex genes suggest that among the XCI escapees, those that are dosage-sensitive genes might have the strongest impact on the phenotype of X0, XXY, and XXX individuals. Importantly, these genes should impact X0, XXY, and XXX individuals in a similar way, as haploinsufficiency or doubled-dosage of protein-complex genes are expected to yield improper stoichiometry in both cases and be deleterious (37). We used the list of protein-complex genes on the X chromosome and identified those escaping XCI (*Materials and Methods*) as likely candidates for X aneuploidy syndromes (Table 1). This list includes the already known *RPS4* Turner candidate gene. The most interesting candidates are probably those involved in large complexes because our results suggest that constraints on dosage are stronger for these. The persistence of a Y homolog also suggests strong constraints on dosage (65) and candidates with a Y homolog and involved in large complexes are in boldface in Table 1. This should not be considered an exhaustive list because the data on X-inactivation and protein-complex genes (and dosage-sensitive genes in general) are known to be partial.

Klinefelter syndrome is characterized by high stature, sparse body hair, gynecomastia, infertility, small testes, decreased verbal intelligence, and increased risks for autoimmune diseases (60). Features of Triple-X syndrome include tall stature, epicanthal folds, hypotonia, clinodactyly, seizures, renal and genitourinary abnormalities, premature ovarian failure, motor and speech delays, and increased risks of cognitive deficits and learning disabilities (59). Turner syndrome is characterized by short stature, premature ovarian failure, and a variety of anatomic abnormalities, including webbing of the neck, lymphedema, aortic coarctation, autoimmune diseases, and characteristic neurocognitive deficits (impaired visual-spatial and visual-perceptual abilities, motor function, nonverbal memory, executive function and attentional abilities; see ref. 63). Interestingly, some of the candidate genes have annotations reminiscent of these X aneuploidy features, although the syndromes are not explicitly cited (Table S1). In addition, three of the four best candidates (in large complexes and with a Y homolog; in boldface in Table 1) are involved in the ubiquitin pathway, which relates to protein degradation and addressing in the cell. Ubiquitination occurs in a wide range of cellular processes, such as differentiation and development, immune response and inflammation, neural and muscular degeneration, morphogenesis of neural networks, and ribosome biogenesis.

Conclusions

Our results open perspectives for finding candidate genes for X aneuploidy syndromes. Such syndromes are very common (up to 1 in 500 male births for Klinefelter) and, although the phenotypic consequences are mild and vary a lot among individuals, with some individuals being asymptomatic, many practicians call for efficient diagnosis because many X aneuploidy individuals can experience health, fertility, and cognitive difficulties if not treated (59, 60, 63). Surprisingly for such common diseases, very little is known about the genotype-phenotype relationships. We suggest dosage-sensitive genes that escape XCI should be tested, for example in animal models (66), as they seem to be good candidate genes for X aneuploidy syndromes.

Our results also show that Ohno's idea of a two-step dosage-compensation mechanism (twofold increase of X expression in both sexes plus an XCI in females) is valid for dosage-sensitive genes (i.e., protein-complex genes). How this two-step dosage-compensation mechanism evolved still needs to be understood. In Ohno's logic, the doubling step should come first and then the halving one (through XCI). However, we know that XCI is very old because the *Xist* locus is located within the earliest diverging segment of the sex chromosomes (stratum 1; see ref. 3) and XCI is found in both marsupials and placentals, which suggests XCI may have evolved first. As the range of XCI silencing crept along the chromosome, then X-linked dosage-sensitive genes (but not other genes) would have experienced selection for doubling of expression. However, in this case, the reason why XCI would evolve first is not clear. In marsupials and in some tissues (placenta, brain) of some placentals, XCI always affects the paternal

Table 1. List of candidate genes for X aneuploidy syndromes

Gene name*	Y homology	Max complex size†	Function annotation‡
PPP2R3B	Y homolog	3	Serine/threonine-protein phosphatase 2A regulatory subunit B
TBL1X	Y homolog	7	F-box-like protein involved in the recruitment of the **ubiquitin**/19S proteasome complex to nuclear receptor-regulated transcription units
RBBP7	-	16	Core histone-binding subunit, Component of several complexes which regulate chromatin metabolism
EIF1AX	Y homolog	3	Eukaryotic translation initiation factor 1A
SH3KBP1	-	7	SH3 domain-containing kinase-binding protein 1
USP9X	Y homolog	20	**Ubiquitin**-specific-processing protease FAF-X.
MED14	Y pseudogene	29	Mediator complex, a coactivator involved in the regulated transcription of nearly all RNA polymerase II-dependent genes
UBA1	Y homolog	40	**Ubiquitin**-activating enzyme E1
WAS	-	12	Effector protein for Rho-type GTPases. Regulates actin filament reorganization via its interaction with the Arp2/3 complex.
SMC1A	-	9	Central component of cohesin complex, required for the cohesion of sister chromatids after DNA replication.
RPS4	Y homolog	40	40S ribosomal protein S4, structural constituent of ribosome
MAGEE1	-	4	Hepatocellular carcinoma-associated protein 1
CHM	-	3	Rab proteins geranylgeranyltransferase component A 1
MORF4L2	-	27	Component of the NuA4 histone acetyltransferase complex
TRPC5	-	5	Transient receptor potential Ca2+ channel
PLS3	-	3	Actin-bundling protein
CUL4B	-	7	Core component of multiple cullin-RING-based E3 **ubiquitin**-protein ligase complexes
HCFC1	-	13	Host cell factor 1

*The best candidates (members of large complexes and with a Y homolog) are shown in bold.
†Most of the genes are involved in several complexes in the list from HPRD (see *Material and Methods*), only the size of the largest complex is indicated here.
‡From NextProt, the new database on human proteins developed by Swissprot (www.nextprot.org). Ubiquitin related genes are in bold.

X (67–71). Some authors suggested that XCI may have originally been a form of genomic imprinting related to parental conflicts (72, 73). Further work is needed to distinguish these two alternatives, but in any case, our work establishes the role of XCI in balancing expression between X and autosomal genes that are dosage-sensitive.

Materials and Methods

Expression Data. We used gene-expression levels obtained from RNA-Seq data of 19,800 human genes (19,066 autosomal and 734 X) in 12 male and female tissues compiled by Xiong et al. (15). Sources of RNA-Seq data and methods are described in ref. 15 but, briefly, only reads uniquely mapped to exons were considered valid hits and expression level of a gene was defined by the number of valid hits to the gene divided by the effective length of the gene. For comparisons between tissues or developmental stages, expression levels were normalized by dividing the total number of valid hits in the sample. Genes with effective length smaller than 100 were discarded, resulting in 19,800 genes.

We cross-linked ref. 15 and ref. 17 datasets using gene names (as no other identifier was available in the latter). We could keep 9,835 genes, which revealed that expression estimates from both studies are strongly correlated: lung (Spearman $\rho = 0.87$), adipose ($\rho = 0.91$), brain ($\rho = 0.92$), colon ($\rho = 0.88$), heart ($\rho = 0.94$), liver ($\rho = 0.94$), lymph node ($\rho = 0.90$), muscle ($\rho = 0.94$), testes ($\rho = 0.85$), kidney ($\rho = 0.89$), breast ($\rho = 0.88$); all with a P value $< 10–5$.

Protein-Complex Data. We obtained a list of members of human protein complex from HPRD release 9 (www.hprd.org). This list includes 1,521 annotated (and experimentally confirmed) protein complexes (74). Human genes and their chromosomal locations (X, autosomal) as described in Ensembl release 52 (www.ensembl.org) were assigned to the members of protein complexes using Ensembl IDs in HPRD. Using protein-complex IDs, we counted the number of the members for each complex to get the protein-complex size. Members without any Ensembl gene IDs were excluded, as well as complexes including only X or autosomal genes. This process led to a dataset of 207 complexes with proteins from 235 X and 1,381 autosomal genes and 89 X and 800 autosomal unique genes, as some genes are involved in several complexes.

X-Inactivation Data. We used data on XCI from ref. 58. These data were obtained constructing nine different rodent/human somatic cell hybrids that retained an inactivated human X. National Center for Biotechnology Information (NCBI) build 34.3 annotations of X genes was used to design primers to amplify mRNAs and quantify X-inactivation of human X genes (58). Using these data, we classified as "inactivated" the genes that were significantly expressed only in two cells or less, as "escaping" the ones for which at least seven cells with a significant expression was observed, and as "heterogeneous" all other genes.

We checked all of the primers by blasting them on the updated X chromosome sequence from Ensembl release 60 (www.biomart.org). From the original 634 genes studied by Carrel and Willard (58), only 495 had both primers that matched both opposite strands and were separated by less than 100 Kb on the Ensembl release 60 X chromosome sequence. In some cases, several genes fell in the interval amplified by the same pair of primers; 69 genes were concerned. We excluded pseudogenes and selected the same gene as in ref. 58 when possible, and picked a gene at random in the interval otherwise. We also checked if primers matched on human autosomal chromosomes or on mouse X chromosome. Five genes had both primers that matched on human autosomes (*EIF2S3*, *TIMM8A*, *SEDL*, *DDX3X*, *GLUD1*) and four genes on the mouse X chromosome (*DUSP21*, *HNRPH2*, *PHF16*, *ABCB7*), and all were withdrawn to avoid false-positives of the RT-PCR experiment. We finally obtained a list of 392 genes. Among these, 55 are escapees, 304 are X-inactivated, and 33 are heterogeneous.

Statistical Analysis. About 15% of X genes are known to escape XCI (58); we did not exclude these genes from the dataset as in ref. 15. Details on analysis are found in the figure legends. All statistical analyses were done using R.

ACKNOWLEDGMENTS. We thank Xionglei He and Jianzhi Zhang, and Di Nguyen and Christine Disteche for sharing with us the Xiong et al. (15) and Deng et al. (17) datasets, respectively; Judith Ross, Hugues Roest Crollius, Erika Kvikstadt, Tristan Lefébure, and Susana Coelho for discussions; and two anonymous referees for their constructive comments. This study was supported by Agence Nationale de la Recherche Grant ANR-08-JCJC-0109 (to G.A.B.M.) and a Science Foundation Ireland grant (to A.M.).

1. Ohno S (1967) *Sex Chromosomes and Sex Linked Genes* (Springer, Berlin).
2. Graves JAM (1995) The origin and function of the mammalian Y chromosome and Y-borne genes—An evolving understanding. *Bioessays* 17:311–320.
3. Lahn BT, Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* 286:964–967.
4. Potrzebowski L, et al. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. *PLoS Biol* 6:e80.
5. Veyrunes F, et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. *Genome Res* 18:965–973.
6. Ross MT, et al. (2005) The DNA sequence of the human X chromosome. *Nature* 434: 325–337.
7. Lemaitre C, et al. (2009) Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biol Evol* 1:56–66.
8. Charlesworth B, Charlesworth D (2000) The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* 355:1563–1572.
9. Potrzebowski L, Vinckenbosch N, Kaessmann H (2010) The emergence of new genes on the young therian X. *Trends Genet* 26:1–4.
10. Zhang YE, Vibranovski MD, Landback P, Marais GA, Long M (2010) Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLoS Biol* 8: pii, e1000494.
11. Lyon MF (1961) Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* 190:372–373.
12. Nguyen DK, Disteche CM (2006) Dosage compensation of the active X chromosome in mammals. *Nat Genet* 38:47–53.
13. Gupta V, et al. (2006) Global analysis of X-chromosome dosage compensation. *J Biol* 5:3.
14. Lin H, et al. (2007) Dosage compensation in the mouse balances up-regulation and silencing of X-linked genes. *PLoS Biol* 5:e326.
15. Xiong Y, et al. (2010) RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat Genet* 42:1043–1047.
16. Casci T (2011) What dosage compensation? *Nat Rev Genet* 12:2.
17. Deng X, et al. (2011) Evidence for compensatory upregulation of expressed X-linked genes in mammals, *Caenorhabditis elegans* and *Drosophila melanogaster*. *Nat Genet* 43:1179–1185.
18. Castagné R, et al. (2011) The choice of the filtering method in microarrays affects the inference regarding dosage compensation of the active X-chromosome. *PLoS ONE* 6: e23956.
19. Kharchenko PV, Xi R, Park PJ (2011) Evidence for dosage compensation between the X chromosome and autosomes in mammals. *Nat Genet* 43:1167–1169, author reply 1171–1172.
20. Lin H, et al. (2011) Relative overexpression of X-linked genes in mouse embryonic stem cells is consistent with Ohno's hypothesis. *Nat Genet* 43:1169–1170, author reply 1171–1172.
21. He X, et al. (2011) Author reply. *Nat Genet* 43:1171–1172.
22. Yildirim E, Sadreyev RI, Pinter SF, Lee JT (2011) X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. *Nat Struct Mol Biol* 19:56–61.
23. Itoh Y, et al. (2007) Dosage compensation is less effective in birds than in mammals. *J Biol* 6(22):2.
24. Ellegren H, et al. (2007) Faced with inequality: Chicken do not have a general dosage compensation of sex-linked genes. *BMC Biol* 5:40.
25. Itoh Y, et al. (2010) Sex bias and dosage compensation in the zebra finch versus chicken genomes: General and specialized patterns among birds. *Genome Res* 20: 512–518.
26. Wolf JB, Bryk J (2011) General lack of global dosage compensation in ZZ/ZW systems? Broadening the perspective with RNA-seq. *BMC Genomics* 12:91.
27. Mank JE, Ellegren H (2009) All dosage compensation is local: Gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. *Heredity (Edinb)* 102:312–320.
28. McQueen HA, Clinton M (2009) Avian sex chromosomes: Dosage compensation matters. *Chromosome Res* 17:687–697.
29. Zha X, et al. (2009) Dosage analysis of Z chromosome genes using microarray in silkworm, *Bombyx mori*. *Insect Biochem Mol Biol* 39:315–321.
30. Walters JR, Hardcastle TJ (2011) Getting a full dose? Reconsidering sex chromosome dosage compensation in the silkworm, *Bombyx mori*. *Genome Biol Evol* 3:491–504.
31. Vicoso B, Bachtrog D (2011) Lack of global dosage compensation in *Schistosoma mansoni*, a female-heterogametic parasite. *Genome Biol Evol* 3:230–235.
32. Deakin JE, Hore TA, Koina E, Marshall Graves JA (2008) The status of dosage compensation in the multiple X chromosomes of the platypus. *PLoS Genet* 4:e1000140.
33. Deakin JE, Chaumeil J, Hore TA, Marshall Graves JA (2009) Unravelling the evolutionary origins of X chromosome inactivation in mammals: Insights from marsupials and monotremes. *Chromosome Res* 17:671–685.
34. Leder EH, et al. (2010) Female-biased expression on the X chromosome as a key step in sex chromosome evolution in threespine sticklebacks. *Mol Biol Evol* 27:1495–1503.
35. Mank JE, Hosken DJ, Wedell N (2011) Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. *Evolution* 65: 2133–2144.
36. Birchler JA, Newton KJ (1981) Modulation of protein levels in chromosomal dosage series of maize: The biochemical basis of aneuploid syndromes. *Genetics* 99:247–266.
37. Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. *Nature* 424:194–197.
38. Edger PP, Pires JC (2009) Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. *Chromosome Res* 17:699–717.
39. Birchler JA, Veitia RA (2010) The gene balance hypothesis: Implications for gene regulation, quantitative traits and evolution. *New Phytol* 186:54–62.
40. Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the *Arabidopsis thaliana* proteome. *Trends Genet* 20:461–464.
41. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell* 16:1679–1691.
42. Freeling M, Thomas BC (2006) Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. *Genome Res* 16:805–814.
43. Aury JM, et al. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature* 444:171–178.
44. Hakes L, Pinney JW, Lovell SC, Oliver SG, Robertson DL (2007) All duplicates are not equal: The difference between small-scale and genome duplication. *Genome Biol* 8: R209.
45. Barker MS, et al. (2008) Multiple paleopolyploidizations during the evolution of the Compositae reveal parallel patterns of duplicate gene retention after millions of years. *Mol Biol Evol* 25:2445–2455.
46. Davis JC, Petrov DA (2005) Do disparate mechanisms of duplication add similar genes to the genome? *Trends Genet* 21:548–551.
47. Maere S, et al. (2005) Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci USA* 102:5454–5459.
48. Liang H, Plazonic KR, Chen J, Li WH, Fernández A (2008) Protein under-wrapping causes dosage sensitivity and decreases gene duplicability. *PLoS Genet* 4:e11.
49. Fernández A, Lynch M (2011) Non-adaptive origins of interactome complexity. *Nature* 474:502–505.
50. Makino T, McLysaght A (2010) Ohnologs in the human genome are dosage balanced and frequently associated with disease. *Proc Natl Acad Sci USA* 107:9270–9274.
51. Kondrashov FA, Koonin EV (2004) A common framework for understanding the origin of genetic dominance and evolutionary fates of gene duplications. *Trends Genet* 20:287–290.
52. Gout JF, Kahn D, Duret L; Paramecium Post-Genomics Consortium (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. *PLoS Genet* 6:e1000944.
53. Wijchers PJ, et al. (2010) Sexual dimorphism in mammalian autosomal gene regulation is determined not only by Sry but by sex chromosome complement as well. *Dev Cell* 19:477–484.
54. Monkhorst K, Jonkers I, Rentmeester E, Grosveld F, Gribnau J (2008) X inactivation counting and choice is a stochastic process: Evidence for involvement of an X-linked activator. *Cell* 132:410–421.
55. Monkhorst K, et al. (2009) The probability to initiate X chromosome inactivation is determined by the X to autosomal ratio and X chromosome specific allelic properties. *PLoS ONE* 4:e5616.
56. Ferguson-Smith MA (1965) Karyotype-phenotype correlations in gonadal dysgenesis and their bearing on the pathogenesis of malformations. *J Med Genet* 2:142–155.
57. Fisher EM, et al. (1990) Homologous ribosomal protein genes on the human X and Y chromosomes: Escape from X inactivation and possible implications for Turner syndrome. *Cell* 63:1205–1218.
58. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* 434:400–404.
59. Tartaglia NR, Howell S, Sutherland A, Wilson R, Wilson L (2010) A review of trisomy X (47, XXX). *Orphanet J Rare Dis* 5:8.
60. Tüttelmann F, Gromoll J (2010) Novel genetic aspects of Klinefelter's syndrome. *Mol Hum Reprod* 16:386–395.
61. Yang F, Babak T, Shendure J, Disteche CM (2010) Global survey of escape from X inactivation by RNA-sequencing in mouse. *Genome Res* 20:614–622.
62. Blaschke RJ, Rappold G (2006) The pseudoautosomal regions, SHOX and disease. *Curr Opin Genet Dev* 16:233–239.
63. Ross J, Roeltgen D, Zinn A (2006) Cognition and the sex chromosomes: Studies in Turner syndrome. *Horm Res* 65:47–56.
64. Geerkens C, Just W, Held KR, Vogel W (1996) Ullrich-Turner syndrome is not caused by haploinsufficiency of RPS4X. *Hum Genet* 97:39–44.
65. Park C, Carrel L, Makova KD (2010) Strong purifying selection at genes escaping X chromosome inactivation. *Mol Biol Evol* 27:2446–2450.
66. Wistuba J (2010) Animal models for Klinefelter's syndrome and their relevance for the clinic. *Mol Hum Reprod* 16:375–385.
67. Heard E (2004) Recent advances in X-chromosome inactivation. *Curr Opin Cell Biol* 16: 247–255.
68. Deakin JE, Chaumeil J, Hore TA, Marshall Graves JA (2009) Unravelling the evolutionary origins of X chromosome inactivation in mammals: Insights from marsupials and monotremes. *Chromosome Res* 17:671–685.
69. Al Nadaf S, et al. (2010) Activity map of the tammar X chromosome shows that marsupial X inactivation is incomplete and escape is stochastic. *Genome Biol* 11:R122.
70. Wang X, Soloway PD, Clark AG (2010) Paternally biased X inactivation in mouse neonatal brain. *Genome Biol* 11:R79.
71. Okamoto I, et al. (2011) Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature* 472:370–374.
72. Haig D (2006) Self-imposed silence: Parental antagonism and the evolution of X-chromosome inactivation. *Evolution* 60:440–447.
73. Engelstädter J, Haig D (2008) Sexual antagonism and the evolution of X chromosome inactivation. *Evolution* 62:2097–2104.
74. Keshava Prasad TS, et al. (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res* 37(Database issue):D767–D772.

Supporting Information

Pessia et al. 10.1073/pnas.1116763109

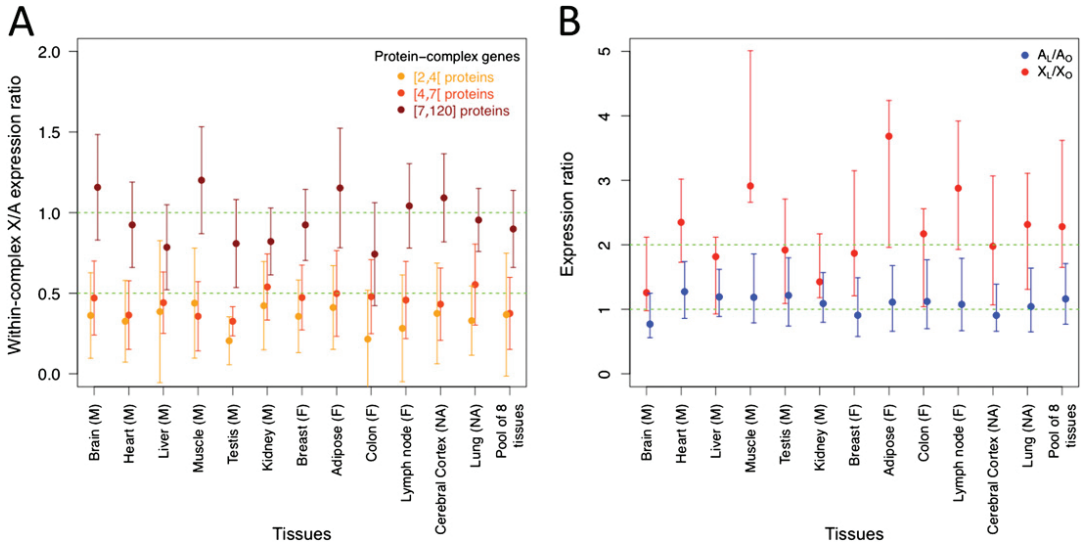


Fig. S1. Protein-complex analysis without nonexpressed genes. (*A*) See Fig. 1*B* legend. (*B*) See Fig. 2 legend. This figure was prepared excluding non-expressed genes.

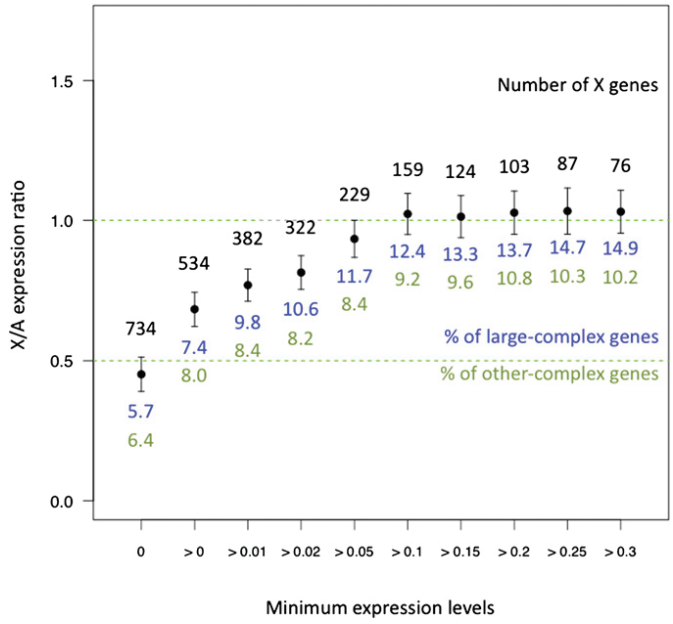


Fig. S2. X/A expression ratio excluding nonexpressed or poorly expressed genes. We show the X/A expression ratio computed as in Fig. 1*A* using pooled expression data for different thresholds for minimum expression levels (from 0 to ≥ 0.3) using Xiong et al.'s (1) expression data. For each threshold, we show the number of remaining X genes and the percentage of large-complex and other-complex genes (as defined in Fig. 2). The two green dashed lines indicate expectations with dosage compensation (X/A = 1) and without dosage compensation (X/A = 0.5).

1. Xiong Y, et al. (2010) RNA sequencing shows no dosage compensation of the active X-chromosome. *Nat Genet* 42:1043–1047.

Table S1. List and description of candidate genes for X aneuploidy syndromes

Gene name[†]	X regions	X-inactivation[‡]	Y homology	Max complex size[§]	Medical annotation[¶]	Function annotation[‖]
PPP2R3B	PAR1	Escaping	Y homolog	3	-	Serine/threonine-protein phosphatase 2A regulatory subunit B
TBL1X	Xp22.2	Escaping	Y homolog	7	Deafness	F-box-like protein involved in the recruitment of the ubiquitin/ 19S proteasome complex to nuclear receptor-regulated transcription units
RBBP7	Xp22.2	Escaping	-	16	-	Core histone-binding subunit, Component of several complexes which regulate chromatin metabolism
EIF1AX	Xp22.12	Escaping	Y homolog	3	-	Eukaryotic translation initiation factor 1A
SH3KBP1	Xp22.12	Heterogeneous	-	7	-	SH3 domain-containing kinase-binding protein 1
USP9X	Xp11.4	Escaping	Y homolog	20	-	Ubiquitin-specific-processing protease FAF-X.
MED14	Xp11.4	Escaping	Y pseudogene	29	-	Mediator complex, a coactivator involved in the regulated transcription of nearly all RNA polymerase II-dependent genes
UBA1	Xp11.23	Escaping	Y homolog	40	Spinal muscular atrophy X-linked type 2 (hypotonia, areflexia, and multiple congenital contractures)*	Ubiquitin-activating enzyme E1
WAS	Xp11.23	Heterogeneous	-	12	Wiskott-Aldrich syndrome (X-linked recessive immunodeficiency characterized by eczema, thrombocytopenia, recurrent infections, and bloody diarrhea)*	Effector protein for Rho-type GTPases. Regulates actin filament reorganization via its interaction with the Arp2/3 complex.
SMC1A	Xp11.22	Escaping	-	9	Cornelia de Lange syndrome X-linked (developmental disorder associated with facial dysmorphisms, abnormal hands and feet, growth delay, cognitive retardation and various other malformations including gastroesophageal dysfunction and cardiac, ophthalmologic and genitourinary anomalies)*	Central component of cohesin complex, required for the cohesion of sister chromatids after DNA replication.
RPS4	Xq13.1	Escaping	Y homolog	40	Turner syndrome candidate	40S ribosomal protein S4, structural constituent of ribosome
MAGEE1	Xq13.3	Heterogeneous	-	4	-	Hepatocellular carcinoma-associated protein 1
CHM	Xq21.2	Escaping	-	3	X-linked Choroideremia (blindness)	Rab proteins geranylgeranyltransferase component A 1
MORF4L2	Xq22.2	Heterogeneous	-	27	-	Component of the NuA4 histone acetyltransferase complex
TRPC5	Xq23	Heterogeneous	-	5	-	Transient receptor potential Ca2+ channel
PLS3	Xq23	Heterogeneous	-	3	-	Actin-bundling protein
CUL4B	Xq24	Heterogeneous	-	7	Mental retardation syndromic X-linked Cabezas type (severe intellectual deficit associated with short stature, craniofacial dysmorphism, small testes, muscle wasting in lower legs, kyphosis, joint hyperextensibility, pes cavus, small feet, and abnormalities of the toes. Additional neurologic manifestations include speech delay and impairment, tremor, seizures, gait ataxia, hyperactivity and decreased attention span)*	Core component of multiple cullin-RING-based E3 ubiquitin-protein ligase complexes
HCFC1	Xq28	Escaping	-	13	-	Host cell factor 1

[†]The best candidates (members of large complexes and with a Y homolog) are shown in bold.
[‡]Escaping, gene always escaping X-inactivation; heterogeneous, gene escaping X-inactivation in some cells (see *Material and Methods*).
[§]Most of the genes are involved in several complexes in the list from HPRD (see *Material and Methods*), only the size of the largest complex is indicated here.
[¶]Medical annotation is mainly from NextProt, the new database on human proteins developed by Swissprot (www.nextprot.org). An asterisk denotes when the annotation shares keywords with Turner, Klinefelter, or Triple-X syndromes.
[‖]Function annotation is also from NextProt. Ubiquitin related genes are highlighted in red.

After the publication of the above article, *Cellular and Molecular Life Sciences* (*CMLS*) journal invited Gabriel Marais to write a review on the evolution of dosage compensation in mammals. Gabriel asked me to join him, and we agreed to include two different parts in this review.

First, in order to clarify the controversy between Ohno's hypothesis followers and opponents, we made a summary of their different results and of the possible reasons for their disagreements. I had a significant contribution to this part.

Second, a more speculative part is included, where we developped an hypothesis on how selection for XCI may have occurred independently of dosage compensation needs. I helped in conceiving this hypothesis, but most of this part of the review was done by Gabriel Marais and Jan Engelstädter.

This paper was sent to *CMLS* for review on the 31^{th} of July 2013, accepted on the 14^{th} of October 2013, and is currently in press.

The evolution of X chromosome inactivation in mammals: the demise of Ohno's hypothesis?

Eugénie Pessia[1] · **Jan Engelstädter**[2] · **Gabriel AB Marais**[1,3]

1 Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, Centre National de la Recherche Scientifique, Villeurbanne F-69622 cedex, France; **2** School of Biological Sciences, The University of Queensland, Brisbane QLD 4072, Australia; **3** Instituto Gulbenkian de Ciência, P-2780-156 Oeiras, Portugal

Abstract Ohno's hypothesis states that dosage compensation in mammals evolved in two steps: a two-fold hyperactivation of the X chromosome in both sexes to compensate for gene losses on the Y chromosome, and silencing of one X (X-chromosome inactivation, XCI) in females to restore optimal dosage. Recent tests of this hypothesis have returned contradictory results. In this review, we explain this ongoing controversy and argue that a novel view on dosage compensation evolution in mammals is starting to emerge. Ohno's hypothesis may be true for a few, dosage-sensitive genes only. If so few genes are compensated, then why has XCI evolved as a chromosome-wide mechanism? This and several other questions raised by the new data in mammals are discussed, and future research directions are proposed.

Keywords Sex chromosomes · Sex determination · Dosage Compensation · Dosage-sensitive genes · Parental antagonism model · RNAseq data

Non-standard abbreviations
XCI X-chromosome inactivation
PAR pseudoautosomal region
rXCI random X-chromosome inactivation
pXCI paternal X-chromosome inactivation
Xi inactivated X chromosome
PolII RNA polymerase II
ZGA zygote genome activation
NGS next-generation sequencing
Ne effective population size
PAM Parental antagonism model
XIC X-inactivation center

Introduction

Since its origin ∼180 MY ago [1,2] the human Y chromosome has lost ∼97% of the genes originally present on that chromosome [3] (Figure 1A). This massive gene loss on the Y resulted in an imbalance of X-linked gene dose in males with their single X chromosome compared to females with two Xs (Figure 1B). Such dosage imbalance was probably deleterious as in humans, dosage imbalance caused by chromosomal aneuploidies (e.g. monosomy, trisomy) of autosomes the size of the X chromosome are lethal. Specific mechanisms have evolved in animals to compensate for this gene dose problem in males, and the solution found in mammals appears to be a complicated one. After Susumu Ohno and Mary Lyon discovered independently female X-chromosome inactivation (XCI) in mammals [4,5], Ohno proposed that dosage compensation in mammals evolved as a two-step mechanism with (1) a two-fold expression increase of the X chromosome in both sexes, which solves the gene dose imbalance problem in males, and (2) inactivation of one of the two X chromosomes by XCI in females to restore optimal dosage [6] (Figure 1B).

XCI has been widely studied both at the mechanistic and evolutionary level (reviewed in e.g. [12,13,14,15,16,17,18,19,20]). In placental mammals, the silencing of the inactivated X (Xi) is established through epigenetic signals involving a long non-coding RNA called *Xist* and a number of cis and trans factors affecting its transcription. In particular, this involves *Rnf12* (main activator of *Xist*), and *Rex1* (main inhibitor of *Xist*) [21]. RNF12 is an E3 ubiquitin ligase that binds REX1, which triggers its degradation by the ubiquitin pathway. This results in *Xist* being activated or not depending on RNF12 dose, which differs between male and female as *Rnf12* is a X-linked gene [21]. XCI is initialized in a particular region of the X chromosome (called the X inactivation center, XIC) where the *Xist* gene is located. It then spreads across the X with *Xist* being transferred directly from XIC to distal sites across the X chromosome that are defined not by specific sequences but by their spatial proximity in the nucleus to XIC [22,23]. XCI is established early during development but

* Corresponding author: gabriel.marais@univ-lyon1.fr

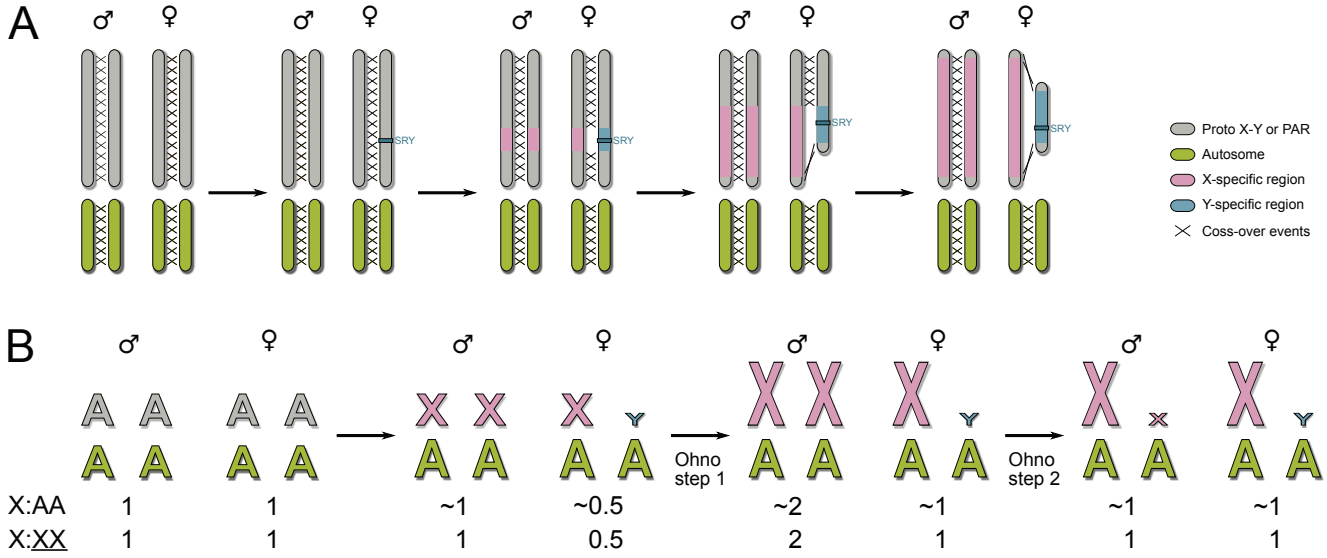


Fig. 1 Sex chromosome and dosage compensation evolution in mammals. (A) Sex chromosome evolution in mammals. *Sry*, the male-determining gene, initiated the evolution of the sex chromosomes from a pair of autosomes. The proto-X and proto-Y stopped recombining at a region including *Sry* and probably other genes, thus forming X- and Y-specific regions. These regions grew larger during evolution through additional recombination suppression events (probably inversions on the Y chromosomes), and gradually diverged. Pseudo-autosomal regions (PARs) are remnant of the autosomal ancestry of the sex chromosomes. In its non-recombining male-specific region the Y has lost most of its genes because of degenerative processes collectively known as Hill-Robertson effects [7,8] **(B)** The evolution of dosage compensation in mammals, as hypothesized by Ohno [6], formalized later by Charlesworth [9,10] and explicitly modelled by [11]. The gene loss on the Y implies dosage imbalance between the sex chromosomes and the autosomes in males. In the first step of Ohno's hypothesis, expression of the X chromosome is doubled in both sexes; proper dosage is restored in males, but is now twice the dosage of autosomes in females. In the second step, the inactivation of one of the two Xs in females evolves in order to get the dosage of the X back to autosomal level. The predicted values of both the expression ratios X:AA and X:XX (see Box 1 and text for more details) are shown at the bottom of panel B. Expression on the autosomes may have changed during evolution, hence the less precise prediction for the X:AA ratio than for X:XX one (Box 1), as emphasized by the '∼' symbol.

there are substantial mechanistic and timing differences among species [18,19]. XCI has evolved region by region on the X chromosome, starting with the region where recombination between X and Y ceased and Y degeneration started first, and encompassing more and more of the X chromosome as recombination suppression progressed [24]. XCI is considered a chromosome-wide phenomenon, but interestingly, 10-15% of the genes on the human X chromosome escape XCI. This includes not only the genes in the still- recombining portion of the sex chromosomes ("pseudoautosomal" regions, PARs), but also blocks of genes in the X-specific region including X-linked genes with a still active Y homolog [24, 16,25]. XCI has probably evolved early in the evolution of the mammalian sex chromosomes, even though *Xist* has emerged only in the placental lineage from a protein-coding gene called *Lnx3* [26].

For a long time, data on the two-fold expression increase on the X chromosome was lacking. Thus, the first step in Ohno's hypothesis has remained very speculative, but at the same time was widely accepted in the scientific community. Only when chromosome-wide analysis of gene expression became possible has Ohno's hypothesis started to be tested. The first studies using microarray seemed to support Ohno's idea of X expression doubling in both sexes [27,28,29,30,31]. However, the first study using RNAseq, a next-generation sequencing technology (NGS) to study gene expression, found no evidence for Ohno's hypothesized first step [32]. An avalanche of papers followed, some supporting Ohno and others contradicting him [33,34,35,36, 37,38,39,40]. The aim of this review is to explain this ongoing controversy, to show that despite the controversy a novel view on dosage compensation evolution in mammals is starting to emerge and to highlight the fundamental questions that remain to be answered.

The controversy about testing Ohno's hypothesis

In mammals, an early study found some support for Ohno's hypothesized X expression doubling by showing that an autosomal gene in laboratory mouse strains exhibited a doubling of its expression following translocation to the X chromosome in *Mus spretus* [42]. However, it was only later with the advance of the microar-

Table 1 Summary of recent studies testing for Ohno's hypothesis

Human X:AA	Mouse X:AA	Study concluded to global hypertranscription	Expression data	Dataset filtering	References
0.9	1	Yes	Microarray	***	[27]
-	1	Yes	Microarray	***	[28]
0.5	0.2	No	RNAseq	* excluding same proportion of lowly and highly expressed genes from A and X	[32]
0.9 **(0.6)**	0.8 **(0.5)**	No if process for filtering changed	Microarray	*** **(same proportion of genes from A and X)**	[41]
0.9	0.9	Yes	RNAseq and PolII occupancy	***	[33]
0.5	-	No	RNAseq	*** grouped by expression levels	[34]
0.9	0.8	Yes	RNAseq	***	[35]
-	1	Yes	Microarray	**	[36]
-	0.8	Yes	RNAseq and PolII occupancy	**	[40]
0.7 **(0.9)**	-	No, except for complexes $\geq$ 7 proteins	RNAseq	** **(from complexes $\geq$ 7 proteins)**	[37]
0.5 / 0.5#	0.4 / 0.5#	No, except hypotranscription for some interaction networks proteins	RNAseq	** conserved in all amniotes	[38]
0.5 / 0.5# **(0.9#)**	0.3 / 0.4#	No, except for complexes $\geq$ 7 proteins	RNAseq	* **(from complexes $\geq$ 7 proteins)**	[39]

() Values given in brackets and in bold correspond to data subset indicated in the Dataset filtering column (also in bold)
X:XX ratio
* All genes
** Expressed Genes (FPKM > 0)
*** Actively Expressed Genes (FPKM $\geq$ 1 or FPKM $\geq$ 3 or detected in $\geq$ 95% of microarray samples)

ray technology that Ohno's hypothesis could be tested with many genes (Table 1). This was done by comparing the global expression on the X chromosome to that on the autosomes and computing the X:AA expression ratio (Box 1). In accord with Ohno's hypothesis, a mean X:AA expression ratio close to one in male as well as female tissues from several mammalian species including human, macaque, mouse and rat was obtained [27,28]. Moreover, similar X expression in both sexes was found in human and mouse. Several other microarray studies were conducted and they all supported Ohno's hypothesis. However, the accuracy of the expression level estimates from microarray data was later criticized (Box 1).

In 2010, the first study using RNAseq, a NGS technology supposed to give much better estimates (Box 1), reported a median X:AA ratio close to 0.5 in a variety of human tissues (both from male and female) and an even lower ratio in mice, challenging Ohno's hypothesis [32]. A year later, several other studies using more RNAseq data were published (Table 1). The 2010 study was criticized for having included genes with no or very low (probably noisy) expression in their analysis. The X chromosome includes more tissue-specific (mostly testis-specific) genes than the autosomes. This means that for instance in liver tissue, many testis-specific genes on the X have no expression and because the X has more of these genes than the autosomes, the X:AA ratio is reduced. When only the genes expressed in one tissue were included to compute the X:AA ratio of that tissue, X:AA ratios were much closer than 1 in both humans and mice [33]. Moreover, ChIP-chip data in mice showed a relatively higher occupancy of an active form of RNA polymerase II (PolII) for highly expressed X-linked genes compared to highly expressed autosomal genes, giving more support to the idea of X hyperexpression proposed by Ohno [33]. Quite strikingly, the same issue of *Nature Genetics* included a total of five articles reporting tests of Ohno's hypothesis, some reporting a X:AA close to 1 [33,35], another reanalyzing microarray data and confirming a X:AA of 1 [36]. A paper published in another journal also reported higher PolII occupancy and more active histone marks

Box 1 Testing Ohno's hypothesis with expression data.
Ohno's hypothesis has been tested by comparing the expression of the X chromosome to that of the autosomes taken together, the X:AA ratio.
Microarray versus RNAseq data Initially, microarray data have been used for this test [27,28,29,30,31]. Microarray may give less precise estimates of expression levels [32]. Moreover, microarray data have to be filtered prior to analysis. The procedures for data filtering rely on arbitrary thresholds, which applied similarly to the X and autosomes remove many lowly expressed X-linked genes and generate an artifactual X:AA of 1 [41]. RNAseq data are supposed to give more precise estimates of expression levels. However, there is also some noise in RNAseq data due to unspecific mapping of RNAseq reads onto the genome, and how this noise is removed can also affect the results [43]. Removing this noise is at the heart of the controversy between the different studies using RNAseq [32,33,34,35,40]. As the threshold for considering a given expression level different from 0 increases, the X:AA ratio increases and reaches a plateau at 1 [33]. It is clear, however, that when using conservative thresholds, the number of X-linked genes analyzed becomes small, and one cannot conclude from this about a "global" X hyperactivation [37].
X:AA, X:XX and other expression ratios Using X:AA expression relies on the assumption that expression were similar between the proto-sex chromosomes and the autosomes (XX:AA = 1). Using the present-day and ancestral expression of the X chromosome, the X:XX ratio, is thus a more direct way to test for Ohno's hypothesis. Computing the X:XX ratio in mammals has implied finding an outgroup where the 1-to-1 orthologs of the X-linked genes are autosomal, namely birds [38]. This guarantees that only genes that were originally on the sex chromosomes before they diverged are analysed, which is what should be done as dosage compensation is expected for these genes only. The new genes that evolved (e.g. through intra-X duplication or translocation to the X) after X and Y stopped recombining and diverged should not be included in studies on dosage compensation, are correctly excluded of the X:XX analysis but not in the X:AA ones. However, finding 1-to-1 orthologs between distantly related species may be difficult and result in a small number of genes being analysed. Moreover, all these chromosome-wide comparisons may be problematic as different selective forces (dosage compensation, sexual selection) may affect expression levels [44,45]. A more precise way of testing Ohno's hypothesis is to study X-linked and autosomal genes that are expected to interact in some ways and for which equal dosage may be required. Considering genes from the same network is one possibility [38], and considering genes belonging to protein complexes is another [37].

on the active X chromosome than on the autosomes [40]. Ohno's hypothesis seemed to regain support.

However, in their reply to these papers, the 2010 paper authors criticized excluding genes with no expression as being arbitrary [34]. They noticed that testing Ohno's hypothesis using X:AA ratios will work only if global expression of the proto-X and the autosomes was initially the same –an assumption that was never tested. As global expression varies even among autosomes, it is possible that the expression of the proto-X differed from that of the other chromosomes even before Y gene loss had started. If this were the case, the present-day X:AA is not expected to be one even if Ohno's hypothesis is correct and expression doubling took place during evolution. A much better way of testing Ohno's hypothesis is to compare the present-day expression of the X chromosome to its ancestral expression, the X:XX ratio (Box 1). This has recently been done by focusing on the genes that have 1:1 orthologs between chicken and humans using an amniote-wide RNAseq dataset [46,38]. Importantly, these "old" genes were initially present on the proto-X chromosome and are those that should show patterns of dosage compensation. By contrast, such patterns are not expected for the many "young" genes gained late in the evolution of the X chromosome [47]. The ancestral expression of the "old" genes was estimated using the expression of their autosomal orthologs in chicken, and the X:XX was found to be 0.5 for placental species including human, chimpanzee, bonobo, gorilla, orangutan, macaque and mouse. In marsupials (opossum), however, the X:XX expression ratio was found to be one. Interestingly, this study also showed that the assumption that expression was similar in the proto-X and the autosomes underlying all the studies using X:AA was actually correct (XX:AA $\approx$ 1 and AA:AA $\approx$ 1). Even measuring the X:AA ratio for the "old" genes returned a value of 0.5 as noted previously [34]. Including or excluding the tissue-specific genes did not change anything in this pattern (as tissue-specific genes are mostly "young" genes). This was later confirmed by an independent analysis of the same dataset, which found a X:XX of 0.5 for all placentals and also marsupials tested [39]. For unknown reasons, the results for marsupials differ in both studies. Importantly, X:AA ratios of 0.5 were confirmed using protein abundance, which suggests that the observed patterns are robust to the method of measuring expression [32,39].

Dosage compensation of a minority of dosage-sensitive genes

The latest tests of Ohno's hypothesis using ancestral X expression seem thus to reject it (Figure 2A), and even the tests using X:AA expression ratio do not fully agree with Ohno's hypothesis (in [33,35,39] X:AA ratios are lower than one when including poorly to moderately as well as highly expressed genes in the analysis, as noted in [37]). Are the sex chromosomes left with dosage problems in mammals? In humans, removing one chromosome is usually lethal. Y degeneration

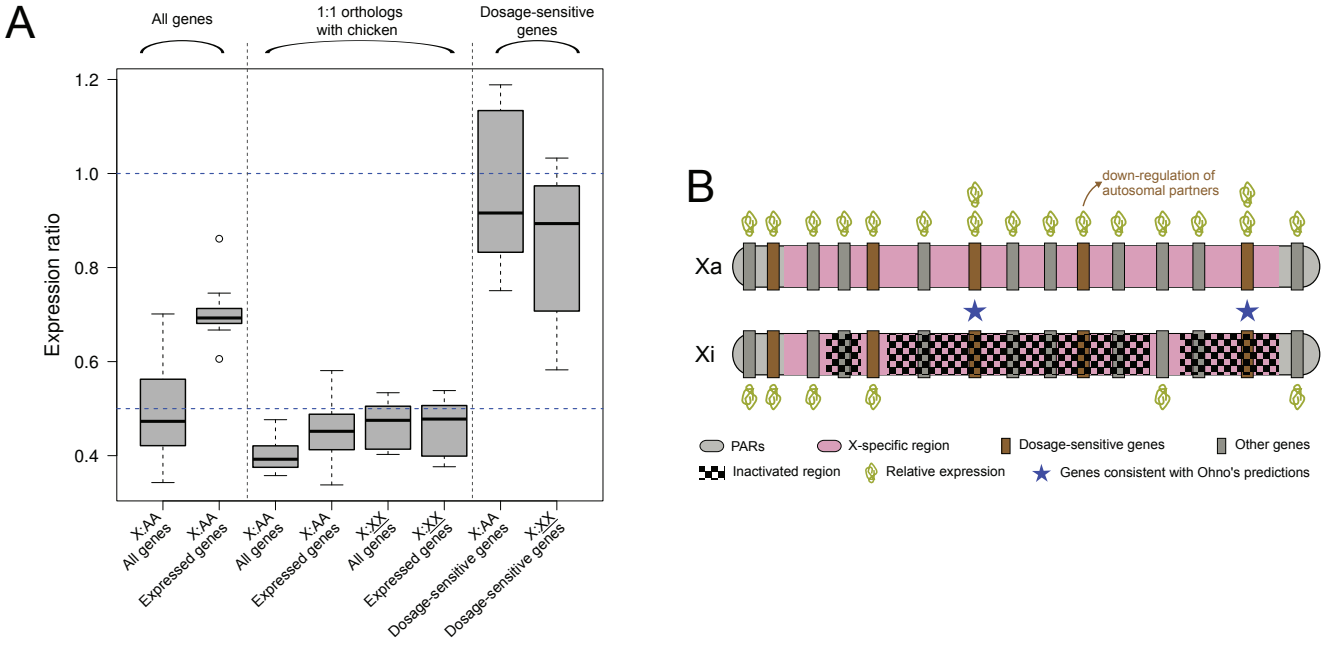


Fig. 2 Dosage compensation of a minority of human X-linked genes. (A) X:AA and X:XX ratios for humans are shown (find more details about these ratios in the text and in Box 1). Results are shown for (1) all human genes or including only genes with a minimum expression level of FPKM >1 (All genes), (2) 1:1 orthologs between human and chicken, for which ancestral expression could be computed using expression data in chicken (Box 1), considering all of them or only those with a minimum expression level of FPKM >1 (1:1 orthologs with chicken), and (3) genes involved in large protein complexes (with 7 or more proteins) that are likely dosage-sensitive (Dosage-sensitive genes). Boxplots of the X:AA or X:XX medians for different tissues are shown. Extreme outliers can be seen for "Expressed genes" ("All genes" category), they correspond to Brain (highest ratio) and lung (lowest ratio). Data for preparing the "All genes" part and the X:AA of "Dosage-sensitive genes" part are from [37] and are based on 12 tissues. All the other boxplots were obtained from 10 tissues in [39]. The blue dashed lines indicate the expected ratios with global dosage compensation (1), and without any dosage compensation (0.5), see text and Figure 1B for more details. **(B)** Sketch summing up the differences in dosage compensation status and mechanisms among the genes on the X chromosome. Most of the genes on the Xi are inactivated, except for the PARs and some XCI-escapees, and XCI is thus a global process. The hyperexpression, on the contrary, appears to be a local process affecting only the dosage-sensitive genes. Dosage compensation through hyperexpression and XCI as envisioned by Ohno thus only affects dosage-sensitive genes [37]. Some dosage-sensitive genes are compensated through another mechanism, namely down-regulation of their autosomal partners, as shown for some genes involved in protein-protein interaction networks [38].

however took millions of years and was gradual, and the dosage problems may not be as severe as in an instantaneous loss of a chromosome. Also, buffering mechanisms that can partially compensate for the loss of a chromosome do exist [48]. Another possibility is that dosage compensation in mammals is a "half-full, half-empty glass" problem, with some genes being compensated but not all [49]. Looking at all the genes at the same time returns a X:AA ratio between 0.5 and 1, which some consider consistent with Ohno's hypothesis (half-full glass) and others inconsistent with the same hypothesis (half-empty glass).

Many genes are known to be haplosufficient or dosage-insensitive [50], i.e. it is not lethal to lose one functional copy of those genes. For instance, it has been recently shown using a theoretical approach that X-linked genes involved in metabolic networks can easily lose a copy without a significant effect on the flux of the network and on fitness, especially for networks with many steps [51]. Dosage compensation should evolve for haploinsufficient or dosage-sensitive genes only [45, 49]. No exhaustive list of mammalian dosage-sensitive genes is available, but there are some known good candidates. The genes encoding proteins involved in complexes (protein-complex genes) are among those candidates. Stoichiometry of the components of a complex is required for its proper folding and functioning [50]. Using human RNAseq data, two of us studied the X:AA expression ratio within complexes, and found that the X:AA for large complexes (seven or more proteins) is around one [37]. Also, it was shown that X expression increase (and not autosomal expression decrease) explains this result by comparing X and autosomal expression among large and small complexes, and finding similar autosomal expression and increased X expression in large versus small complexes [37]. This was later confirmed by using ancestral expression levels and looking at each complex's X:XX and AA:AA

ratio [39]. Other dosage-sensitive candidates are genes involved in protein-protein interaction networks. Julien et al. (2012) studied those genes and found that in some cases, autosomal genes have evolved reduced expression following their X-linked partners, another way of achieving dosage compensation [38]. Dosage-sensitive genes on the X chromosome are thus dosage- compensated, some showing the hyperexpression proposed by Ohno [52] (Figure 2, Table 1). However, these genes represent a minority of the X-linked genes (even though there are probably some unidentified ones on the X [53]), which explains the "half-full, half-empty glass" problem when studying Ohno's hypothesis with all X-linked genes.

The mechanisms ensuring dosage-sensitive genes to match the expression of their autosomal partners in complexes or in protein networks are unknown. In some cases, expression of the X-linked genes was increased [37,39], in other cases expression of the autosomal genes was decreased [38], which suggests that these mechanisms evolved on a gene-by-gene basis. Using ChIP-chip or ChIP-seq in mice, a higher RNA Polymerase II occupancy was found on the X chromosome than on the autosomes [33,40]. More epigenetic marks characteristic of actively transcribed genes were found on the X chromosome compared to autosomes [40]. Interestingly, these trends were only found for highly expressed genes (also noted in [49]), of which dosage-sensitive genes may represent a substantial fraction [54,55,37]. However, it would be important to explicitly compare epigenetic patterns of dosage-sensitive versus other genes in order to gain a better understanding of the different mechanisms of adjusting expression of dosage-sensitive genes.

The origin of XCI and the early steps of dosage compensation evolution

Only a minority of genes show clear evidence of dosage compensation in mammals. If XCI initially evolved to counteract hyperactivation of the X chromosome in females as proposed by Ohno, it is not clear why XCI is global and affects the majority of the X-linked genes when hyperactivation is local and affects only a few genes. Of course, XCI may be global for unknown mechanistic reasons and its effect on many haplosufficient / dosage-insensitive X-linked genes may simply be neutral. It is also possible that XCI evolved for a completely different reason in the first place, and was only later in evolution recruited for dosage compensation (Figure 3). This idea is somewhat supported by a population genetic modelling that has shown that XCI can evolve under Ohno's hypothesis under quite restricted conditions only [11].

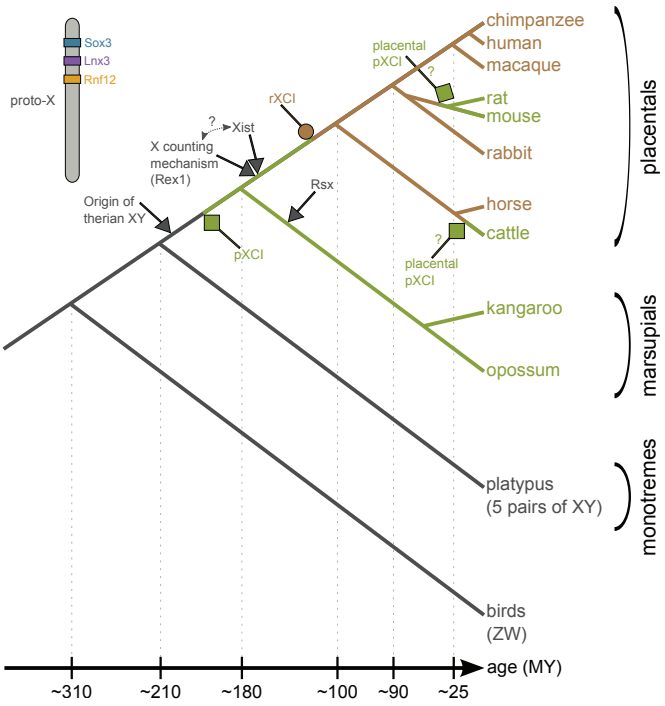


Fig. 3 Steps in the evolution of XCI. Major events the evolution of XCI in placentals and marsupials are shown in the tree of amniotes. XCI, probably pXCI (shown in green), evolved early in the evolution of XY chromosomes, as shown here. However, independent XCI evolution in placental and marsupial lineages cannot be ruled out (see Text). Also shown is how the mechanism of XCI was later refined independently in placentals with the evolution of the lncRNA *Xist* (from the protein-coding gene *Lnx3*) and *Rex1* and became rXCI (shown in brown), and in marsupials with the evolution of the lncRNA with *Xist*-like properties *Rsx* (which has not evolved from *Lnx3*). In some placentals, pXCI is found as well (in early embryo and extra-embryonic tissues); it is not known whether pXCI has re-evolved or has been conserved, hence the question marks. Two major players in the evolution of XCI, *Lnx3* (parent of *Xist*) and *Rnf12*, were probably close to *Sox3* (parent of *Sry*) in the proto-X, and the same small region has apparently been involved in both the evolution of sex determination and XCI (see Text). Birds and monotremes sex chromosome systems evolved independently from that in therians, and serve as outgroups here.

Such an alternative theory for the evolution of XCI was proposed by Haig [56,57]. According to his "parental antagonism model", XCI initially evolved not as a means of dosage compensation, but as a silencing mechanism of growth-inhibiting genes on the X chromosome during embryonic growth (see Box 2 for details). The premise of this hypothesis is the theoretical expectation that the X chromosome is enriched for growth inhibiting genes [58]. Whether or not this premise holds is still unknown, but several lines of evidence are consistent with this idea. It has been shown that the number of X chromosomes affect the speed of early development before XCI is established, with X0 and XY mouse embryos de-

veloping faster than XX embryos (reviewed in [20]). In marsupials, *H19X*, a long coding RNA involved in regulating placenta growth has been found next to *Rsx*, the long coding RNA mediating XCI, which suggests a possible link between imprinting, placenta and XCI [59]. More work is needed but it is possible that the imprinted region was initially large enough for a mechanism such as XCI affecting several neighbour genes at once to evolve, paving the way for global XCI as we know it today. Another attractive feature of Haig's hypothesis is that it predicts imprinted paternal X inactivation (pXCI) rather than random X inactivation (rXCI) to be the primary form of XCI. pXCI is found in marsupials, and also in extra-embryonic tissues and the early developing embryo of some placentals [12,18]. Overall the observations fit with the idea that pXCI was ancestral to rXCI (see next section for more details). However, parallel evolution of pXCI in marsupials and some placentals cannot be ruled out at this point. The parental antagonism model also predicts that XCI should have evolved in groups where parental conflicts about maternal resource allocation to embryonic growth are strong. This is indeed the case in placental and marsupial mammals. Recent data suggest that chromosome inactivation might also affect the bird Z chromosome and the monotreme X chromosomes where parental conflicts may be weaker [60], but this observation needs to be confirmed before taken as evidence against the parental antagonism model.

Another possibility is that the origin of XCI is connected to sex determination [17]. *Sox3*, the gene that gave rise to *Sry*, induces the development of testis when overexpressed, which suggests that evolving appropriate dosage of *Sox3* may have been a crucial point in establishing sex determination through the *Sox3*/*Sry* gene pair in mammals [17]. The initial function of XCI may have been to reinforce differences in dosage of *Sox3*/*Sry* of XX and XY individuals and ensure that they develop into females and males, respectively. Intriguingly, two key genes for the evolution of XCI, *Rnf12* and *Lnx3*, were probably physically close to *Sox3* in the proto-X chromosome ∼180 MYA, as suggested by the analysis of the location of these genes in mammals and birds [17]. Both the evolution of sex determination and XCI apparently involved the same relatively small region on the proto-sex chromosomes. The mechanism of XCI at the time must have been different from what it is today, as some key players such as *Xist* and *Rex1* evolved later in the placental lineage [26,21]. However, *Lnx3* is involved in the ubiquitin pathway as is *Rnf12* (see Introduction), which suggests that this pathway may have had a critical role in the early evolution of XCI. This hypothesis predicts that if XCI has evolved

Box 2 The parental antagonism model of X chromosome inactivation.

The parental antagonism model (PAM) for the evolution of XCI was proposed by Haig [56,57]. It is embedded within the general evolutionary theory of parental investment in offspring [61] and closely related to the kinship theory of genomic imprinting [56,62]. The argument can be presented in a number of steps.

Step 0 A prerequisite for PAM to work is that offspring are provisioned with an adjustable amount of resources from their mother following fertilization. This is indeed the case in therians where resources are provided through the placenta during embryonic development.

Step 1 At the core of PAM is the expectation that there will be an evolutionary conflict between maternally and paternally derived genes within a developing organism with respect to the amount of resources provided by the mother of that individual. Both genes derived from the mother and from the father will be selected to induce the mother to provide resources. However, the optimal amount of resources provided may be greater for paternally than for maternally derived genes. This is because when females mate with multiple males during their lifetimes, paternal interests will be limited to the current offspring whereas maternal interests extend to all future offspring that a mother will have.

Step 2 The X chromosome is two thirds of the time inherited from the mother but only one third of the time from the father (simply because females have two Xs and males just one). As a consequence, genes on the X chromosome are expected to reflect maternal interests more than paternal ones. In particular, it is expected that genes coding for embryonic growth inhibitors will accumulate on the X chromosome, whereas growth enhancers will be scarce [58].

Step 3 As an evolutionary response to this accumulation of growth inhibitor genes on the X chromosome, there will be selection on paternally inherited genes on the X chromosome to inactivate these genes in embryos, thereby increasing embryo growth. This inactivation may then also spread to other genes on the paternally derived X, either for mechanistic reasons or for dosage compensation. The resulting state of inactivation of the paternally derived X (pXCI) is found in marsupials.

Step 4 pXCI entails that an organism becomes functionally haploid, so that recessive deleterious mutations on the maternally derived X chromosome will be expressed and reduce fitness. This may create selection pressure for random XCI (rXCI), alleviating this burden because half of the cells will then express the functional gene copy [57]. This transition from pXCI to rXCI does not involve parental conflict because the choice of which X chromosome is inactivated does not affect gene dosage.

Step 5 Nevertheless, parental conflict over which of the X chromosomes is inactivated may persist or re-emerge. This is because there may still be imprinted growth inhibitor genes on the X chromosome that are silenced when paternally inherited, so that the maternally derived X chromosome will be under selection to remain the active X. As a consequence, pXCI can re-evolve from rXCI, which may explain pXCI in mouse trophoblast tissues.

to silence *Sox3* in females, *Sox3* must be among the X-inactivated genes, which indeed appears to be the case in mice [63,64]. At this point, this hypothesis is very speculative, but it has an interesting implication. Theory predicts that the early suppression of recombination between proto-X and proto-Y chromosomes must involve either at least two sex-determining genes [65], or a sex-determining gene and at least one sexually antagonistic (beneficial for one sex, harmful for the other) gene [8]. Selection will then favour suppressed recombination between sex-determining genes so that the male-determining alleles remain linked on the Y, and the female-determining alleles remain linked on the X. When a male-beneficial-female-detrimental gene appears on the sex chromosomes, selection will also favour it to be genetically linked to the male-determining locus [66]. In mammals, only one sex-determining gene, *Sry*, has been described, and it has been suggested that sexually antagonistic mutations may have accumulated in the vicinity of *Sry* very early in the evolution of the mammalian sex chromosomes, driving suppression of recombination between the X and Y [8]. If genes other than *Sry* were involved in the early evolution of sex determination of mammals as we suggest, this might be sufficient to explain early suppression of X-Y recombination, and sexually antagonistic genes might only have had a role in later in further suppressing recombination along the sex chromosomes [67]. However, more work is needed to test this hypothesis of multiple sex-determining genes in mammals.

The evolution of random XCI

XCI is found both in placentals and marsupials, consistent with an early evolution of this mechanism soon after the therian sex chromosomes originated ∼180 MY ago. However, the mechanisms of XCI are different in both lineages [12,13,14]. The coating of the future inactivated X is mediated by a long non-coding RNA in both placentals and marsupials, but in placentals this RNA is coded by *Xist* (a gene that evolved specifically in the placental lineage from a protein coding gene; [26]) whereas in marsupials this RNA is coded by the non-orthologous gene Rsx [68]. This may indicate a parallel refinement of the mechanism for XCI independently in both lineages, or simply that XCI originated twice. Moreover, in placentals, XCI is random (rXCI), i.e. one of the two Xs is randomly inactivated in different cells during development (stage E.4.5 in mice when cells start to differentiate), which results in a mosaic of cells with differently inactivated Xs in somatic tissues. In marsupials, XCI always affects the X chromosome transmitted by the father, and is called paternal XCI (pXCI) [69]. In placentals, pXCI has been reported in extraembryonic tissues, and in embryonic tissue early in development prior to rXCI (reviewed in [18,19,20]). There are differences however among placentals. In mouse and cattle, pXCI has been observed in trophoblast tissues [18]. In other placentals studied thus far, no pXCI in has been reported in the extraembryonic tissues (human: [70], rhesus macaque: [71], rabbit: [70], and horse: [72]). A biased inactivation (inactivation of the paternal X was found more frequent than that of the maternal X) has also been reported in neonatal brain in mice, but the bias was small and interpreted as a residual of pXCI [73].

Explaining these differences in XCI between species is challenging. The current view is that pXCI is the ancestral form (Figure 3; see also above), but why pXCI would have evolved first is not clear. One reason could be that evolving pXCI is easier from a mechanistic point of view. In line with this, pXCI appears to require fewer cis and trans factors than rXCI (a 250 Kb *Xist*-transgene is enough to recapitulate pXCI, where rXCI requires a 460 Kb one [20]). In particular, rXCI requires a counting mechanism to inactivate only one X and not both Xs, which is mediated by *Rnf12*/*Rex1* [21]. pXCI does not require a counting mechanism as it is always the X from the father that is inactivated [12], and consistently *Rex1* is absent in marsupials [21]. As argued in the previous section, according to the parental antagonism model of XCI there are also evolutionary reasons why pXCI is expected to evolve first.

Understanding why rXCI would evolve to supersede pXCI is also challenging. With pXCI, females are effectively haploid for the X chromosome and recessive deleterious mutations on the X will be expressed. rXCI will generate tissues made of mosaics of cells in which, overall, both alleles will be expressed. An obvious consequence of rXCI is thus restoring partially diploidy for the X chromosome. The typical example is red-green colour blindness in humans. Red-green colour blindness is due to a recessive deleterious mutation on a X-linked opsin gene. This condition affects mostly males as they will always express the deleterious mutation if present on their single X chromosome. Only females homozygous for this deleterious mutation in the X- linked opsin gene will be in a similar situation. In heterozygous females, the X-linked opsin gene is randomly X-inactivated. A sufficient number of cone cells express the functional allele and can sense colour so that most heterozygous female will not be colour-blind, which explains why red-green colour blindness is much more common in men than women. Recent theoretical work has explored the conditions under which rXCI and pXCI may evolve [74]. If the alleles delete-

rious for female fitness are mostly recessive then rXCI is expected to evolve. If many sexually antagonistic alleles (beneficial for one sex, harmful for the other) are segregating in the population, rXCI is not expected to evolve [74]. For instance, alleles beneficial for males and deleterious for females will generate selection for pXCI [74]. However, it is not clear why paternal XCI should evolve rather than maternal XCI. A higher mutation rate in male than in females (male-biased mutation) would imply that paternally-inherited X chromosome will carry more deleterious mutations, and pXCI should be selected in that case [74], although the effect might be too weak and needs to be studied in more details. A stronger sexual selection in males will also favour pXCI, as the paternally-inherited X chromosome will tend to include sexually-antagonistic genes harmful to females [74]. Connallon & Clark claimed that sexual dimorphism and potentially sexual selection might be overall stronger in marsupials than in placentals, which this may explain why pXCI has been maintained in former ([74] and references therein). However, if the strength of sexual selection is the only driver of the transition from pXCI to rXCI, it is difficult to understand why it has not evolved in strongly sexually dimorphic placental species, such as cervids, pinnipeds and some primates. Alternatively, the parental antagonism model states that pXCI in some placentals may have been conserved or re-evolved as this tissue directly mediates demand from the offspring to the mother [57]. The intensity of parental conflicts for offspring demand may explain why some species have pXCI and others do not but this is purely speculative [57]. Differences in the timing of zygote genome activation (ZGA), the process by which the genome of the zygote starts being expressed, may also explain why pXCI is found in some placentals and not others. In mice, ZGA occurs early and may require a fast way of achieving XCI, hence the presence of pXCI in early mouse development. In other placentals (e.g. humans, rabbits), ZGA occurs later and there is sufficient time for rXCI to be established [70]. However, the timing of ZGA does not correlate very well with the presence/absence of pXCI in placentals, although XCI has been studied in only a few placentals thus far [18].

Concluding remarks and future directions

During the last fifty years, Ohno's hypothesis was taken for granted as it was difficult to see how XCI would have evolved otherwise. NGS data have completely changed our view of the evolution of XCI and dosage compensation in mammals, although the situation may again change when more data become available. Global dosage compensation that was once thought to be a paradigm seems now an exception rather than a rule, as many cases of partial dosage compensation have been reported in birds, fish, and some invertebrates (reviewed in [45]). Global dosage compensation is only confirmed for a handful of species such as *Drosophila* and *C. elegans*. If the evolution of dosage compensation is driven by dosage-sensitive genes as suggested by the results in mammals, we expect it to be partial and affect only a few genes on the X or Z chromosomes. The number of dosage-sensitive genes may also vary from one organism to another, and in species with many such genes on the sex chromosomes, a global dosage compensation mechanism may evolve. In species with large effective population size (*Ne*) such as *Drosophila* and *C. elegans*, selection is very efficient so that the number of genes in the genome that are effectively dosage-sensitive is increased. Such a relationship between *Ne* and the extent of dosage compensation remains to be investigated. Also, identifying dosage-sensitive genes in many species will be crucial to test many of the ideas outlined here. Mammalian species with new sex chromosome pairs (as in some rodents, see [75] and references therein) may be particularly interesting as they raise the question of how the dosage-sensitive genes on the former X cope without XCI, if they indeed lose XCI. Understanding why XCI has evolved in the first place is still a great challenge, as testing the different available hypotheses is not easy. Theoretical work is certainly needed to explore further the different hypotheses. In particular, the parental antagonism model has not been modelled formally yet. Understanding the transition from pXCI to rXCI will probably require systemic surveys of XCI in mammals, as only a few species have been studied thus far. Approaches using NGS may facilitate this task and provide a broad picture of the evolution of XCI in mammals in the near future.

Acknowledgements We thank Fangqin Lin for providing us with the exact X:AA and X:XX median values from [39]. GABM is supported by Agence Nationale de la Recherche (grant ref. ANR-12-BSV7-0002). GABM thanks Instituto Gulbenkian de Ciência for hosting him during several periods strongly overlapping with the writing of this article.

References

1. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res 18: 965–973.
2. Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, et al. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 6: e80.

3. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.
4. Ohno S S, Kaplan WD, Kinosita R (1959) Formation of the sex chromatin by a single x-chromosome in liver cells of rattus norvegicus. Exp Cell Res 18: 415–418.
5. Lyon MF (1961) Gene action in the x-chromosome of the mouse (mus musculus l.). Nature 190: 372–373.
6. Ohno S, et al. (1967) Sex chromosomes and sex-linked genes.(Monographs on endocrinology, Vol. 1.). Berlin, Heidelberg, New York: Springer Verlag. URL `http://www.cabdirect.org/abstracts/19680100985.html`.
7. Charlesworth B, Charlesworth D (2000) The degeneration of y chromosomes. Philos Trans R Soc Lond B Biol Sci 355: 1563–1572.
8. Bachtrog D (2013) Y-chromosome evolution: emerging insights into processes of y-chromosome degeneration. Nat Rev Genet 14: 113–124.
9. Charlesworth B (1978) Model for evolution of y chromosomes and dosage compensation. Proc Natl Acad Sci U S A 75: 5618–5622.
10. Charlesworth B (1996) The evolution of chromosomal sex determination and dosage compensation. Curr Biol 6: 149–162.
11. Engelstädter J, Haig D (2008) Sexual antagonism and the evolution of x chromosome inactivation. Evolution 62: 2097–2104.
12. Deakin JE, Chaumeil J, Hore TA, Marshall Graves JA (2009) Unravelling the evolutionary origins of x chromosome inactivation in mammals: insights from marsupials and monotremes. Chromosome Res 17: 671–685.
13. Lee JT (2011) Gracefully ageing at 50, x-chromosome inactivation becomes a paradigm for rna and chromatin control. Nat Rev Mol Cell Biol 12: 815–826.
14. Livernois AM, Graves JAM, Waters PD (2012) The origin and evolution of vertebrate sex chromosomes and dosage compensation. Heredity (Edinb) 108: 50–58.
15. Jeon Y, Sarma K, Lee JT (2012) New and xisting regulatory mechanisms of x chromosome inactivation. Curr Opin Genet Dev 22: 62–71.
16. Disteche CM (2012) Dosage compensation of the sex chromosomes. Annu Rev Genet 46: 537–560.
17. Gribnau J, Grootegoed JA (2012) Origin and evolution of x chromosome inactivation. Curr Opin Cell Biol 24: 397–404.
18. Dupont C, Gribnau J (2013) Different flavors of x-chromosome inactivation in mammals. Curr Opin Cell Biol 25: 314–321.
19. Ohhata T, Wutz A (2013) Reactivation of the inactive x chromosome in development and reprogramming. Cell Mol Life Sci 70: 2443–2461.
20. Schulz EG, Heard E (2013) Role and control of x chromosome dosage in mammalian development. Curr Opin Genet Dev 23: 109–115.
21. Gontan C, Achame EM, Demmers J, Barakat TS, Rentmeester E, et al. (2012) Rnf12 initiates x-chromosome inactivation by targeting rex1 for degradation. Nature 485: 386–390.
22. Chow JC, Ciaudo C, Fazzari MJ, Mise N, Servant N, et al. (2010) Line-1 activity in facultative heterochromatin formation during x chromosome inactivation. Cell 141: 956–969.
23. Engreitz JM, Pandya-Jones A, McDonel P, Shishkin A, Sirokman K, et al. (2013) The xist lncrna exploits three-dimensional genome architecture to spread across the x chromosome. Science 341: 1237973.
24. Carrel L, Willard HF (2005) X-inactivation profile reveals extensive variability in x-linked gene expression in females. Nature 434: 400–404.
25. Zhang Y, Castillo-Morales A, Jiang M, Zhu Y, Hu L, et al. (2013) Genes that escape x-inactivation in humans have high intraspecific variability in expression, are associated with mental impairment but are not slow evolving. Mol Biol Evol .
26. Duret L, Chureau C, Samain S, Weissenbach J, Avner P (2006) The xist rna gene evolved in eutherians by pseudogenization of a protein-coding gene. Science 312: 1653–1655.
27. Nguyen DK, Disteche CM (2006) Dosage compensation of the active x chromosome in mammals. Nat Genet 38: 47–53.
28. Gupta V, Parisi M, Sturgill D, Nuttall R, Doctolero M, et al. (2006) Global analysis of x-chromosome dosage compensation. J Biol 5: 3.
29. Talebizadeh Z, Simon SD, Butler MG (2006) X chromosome gene expression in human tissues: male and female comparisons. Genomics 88: 675–681.
30. Lin H, Gupta V, Vermilyea MD, Falciani F, Lee JT, et al. (2007) Dosage compensation in the mouse balances up-regulation and silencing of x-linked genes. PLoS Biol 5: e326.
31. Johnston CM, Lovell FL, Leongamornlert DA, Stranger BE, Dermitzakis ET, et al. (2008) Large-scale population study of human cell lines indicates that dosage compensation is virtually complete. PLoS Genet 4: e9.
32. Xiong Y, Chen X, Chen Z, Wang X, Shi S, et al. (2010) Rna sequencing shows no dosage compensation of the active x-chromosome. Nat Genet 42: 1043–1047.
33. Deng X, Hiatt JB, Nguyen DK, Ercan S, Sturgill D, et al. (2011) Evidence for compensatory upregulation of expressed x-linked genes in mammals, caenorhabditis elegans and drosophila melanogaster. Nat Genet 43: 1179–1185.
34. He X, Chen X, Xiong Y, Chen Z, Wang X, et al. (2011) He et al. reply. Nature Genetics 43: 1171–1172.
35. Kharchenko PV, Xi R, Park PJ (2011) Evidence for dosage compensation between the x chromosome and autosomes in mammals. Nat Genet 43: 1167–9; author reply 1171-2.
36. Lin H, Halsall JA, Antczak P, O'Neill LP, Falciani F, et al. (2011) Relative overexpression of x-linked genes in mouse embryonic stem cells is consistent with ohno's hypothesis. Nat Genet 43: 1169–70; author reply 1171-2.
37. Pessia E, Makino T, Bailly-Bechet M, McLysaght A, Marais GAB (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.
38. Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, et al. (2012) Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol 10: e1001328.
39. Lin F, Xing K, Zhang J, He X (2012) Expression reduction in mammalian x chromosome evolution refutes ohno's hypothesis of dosage compensation. Proc Natl Acad Sci U S A 109: 11752–11757.
40. Yildirim E, Sadreyev RI, Pinter SF, Lee JT (2012) X-chromosome hyperactivation in mammals via nonlinear relationships between chromatin states and transcription. Nat Struct Mol Biol 19: 56–61.
41. Castagné R, Rotival M, Zeller T, Wild PS, Truong V, et al. (2011) The choice of the filtering method in microarrays affects the inference regarding dosage compensation of the active x-chromosome. PLoS One 6: e23956.

42. Adler DA, Rugarli EI, Lingenfelter PA, Tsuchiya K, Poslinski D, et al. (1997) Evidence of evolutionary upregulation of the single active x chromosome in mammals based on clc4 expression levels in mus spretus and mus musculus. Proc Natl Acad Sci U S A 94: 9244–9248.
43. Jue NK, Murphy MB, Kasowitz SD, Qureshi SM, Obergfell CJ, et al. (2013) Determination of dosage compensation of the mammalian x chromosome by rna-seq is dependent on analytical approach. BMC Genomics 14: 150.
44. Mank JE, Ellegren H (2009) Sex bias in gene expression is not the same as dosage compensation. Heredity (Edinb) 103: 434.
45. Mank JE, Hosken DJ, Wedell N (2011) Some inconvenient truths about sex chromosome dosage compensation and the potential role of sexual conflict. Evolution 65: 2133–2144.
46. Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, et al. (2011) The evolution of gene expression levels in mammalian organs. Nature 478: 343–348.
47. Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M (2010) Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the x chromosome. PLoS Biol 8.
48. Malone JH, Cho DY, Mattiuzzo NR, Artieri CG, Jiang L, et al. (2012) Mediation of drosophila autosomal dosage effects and compensation by network interactions. Genome Biol 13: r28.
49. Birchler JA (2012) Claims and counterclaims of x-chromosome compensation. Nat Struct Mol Biol 19: 3–5.
50. Papp B, Pál C, Hurst LD (2003) Dosage sensitivity and the evolution of gene families in yeast. Nature 424: 194–197.
51. Hall DW, Wayne ML (2013) Ohno's "peril of hemizygosity" revisited: gene loss, dosage compensation, and mutation. Genome Biol Evol 5: 1–15.
52. Wright AE, Mank JE (2012) Battle of the sexes: conflict over dosage-sensitive genes and the origin of x chromosome inactivation. Proc Natl Acad Sci U S A 109: 5144–5145.
53. Veitia RA (2005) Gene dosage balance: deletions, duplications and dominance. Trends Genet 21: 33–35.
54. Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, et al. (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169: 1915–1925.
55. Gout JF, Kahn D, Duret L, Consortium PPG (2010) The relationship among gene expression, the evolution of gene dosage, and the rate of protein evolution. PLoS Genet 6: e1000944.
56. Haig D (2000) The kinship theory of genomic imprinting. Annual Review of Ecology and Systematics 31: 9–32.
57. Haig D (2006) Self-imposed silence: parental antagonism and the evolution of x-chromosome inactivation. Evolution 60: 440–447.
58. Haig D (2006) Intragenomic politics. Cytogenet Genome Res 113: 68–74.
59. Necsulea A, Soumillon M, Liechti A, Daish T, Baker J, et al. Functionality and evolution of lncrna repertoires and expression patterns in tetrapods. Nature in press.
60. Livernois AM, Waters SA, Deakin JE, Marshall Graves JA, Waters PD (2013) Independent evolution of transcriptional inactivation on sex chromosomes in birds and mammals. PLoS Genet 9: e1003635.
61. Trivers R (1972) Parental investment and sexual selection. (Sexual Selection and the Descent of Man 1871-1971). Campbell B.
62. Haig D (2002) Genomic imprinting and kinship. Rutgers University Press.
63. Collignon J, Sockanathan S, Hacker A, Cohen-Tannoudji M, Norris D, et al. (1996) A comparison of the properties of sox-3 with sry and two related genes, sox-1 and sox-2. Development 122: 509–520.
64. Splinter E, de Wit E, Nora EP, Klous P, van de Werken HJG, et al. (2011) The inactive x chromosome adopts a unique three-dimensional conformation that is dependent on xist rna. Genes Dev 25: 1371–1383.
65. Charlesworth B, Charlesworth D (1978) A model for the evolution of dioecy and gynodioecy. American naturalist 112: 975–997.
66. Rice WR (1987) The accumulation of sexually antagonistic genes as a selective agent promoting the evolution of reduced recombination between primitive sex chromosomes. Evolution 41: 911–914.
67. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb) 95: 118–128.
68. Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, et al. (2012) Rsx is a metatherian rna with xist-like properties in x-chromosome inactivation. Nature 487: 254–258.
69. Wang X, Douglas KC, Vandeberg JL, Clark A, Samollow PB (2013) Chromosome-wide profiling of x-chromosome inactivation and epigenetic states in fetal brain and placenta of the opossum, monodelphis domestica. Genome Res .
70. Okamoto I, Patrat C, Thépot D, Peynot N, Fauque P, et al. (2011) Eutherian mammals use diverse strategies to initiate x-chromosome inactivation during development. Nature 472: 370–374.
71. Tachibana M, Ma H, Sparman ML, Lee HS, Ramsey CM, et al. (2012) X-chromosome inactivation in monkey embryos and pluripotent stem cells. Dev Biol 371: 146–155.
72. Wang X, Miller DC, Clark AG, Antczak DF (2012) Random x inactivation in the mule and horse placenta. Genome Res 22: 1855–1863.
73. Wang X, Soloway PD, Clark AG (2010) Paternally biased x inactivation in mouse neonatal brain. Genome Biol 11: R79.
74. Connallon T, Clark AG (2013) Sex-differential selection and the evolution of x inactivation strategies. PLoS Genet 9: e1003440.
75. Veyrunes F, Chevret P, Catalan J, Castiglia R, Watson J, et al. (2010) A novel sex determination system in a close relative of the house mouse. Proc Biol Sci 277: 1049–1056.

2

Another help from the X? The evolution of X-Y gene conversion in primates

Just before my PhD, a collaboration between my lab in Lyon and the lab of Brigitte Crouau-Roy in Toulouse (France) was initiated. For two years, Toulouse designed lots of primers allowing to sequence five X-linked genes and their five Y-linked gametologs in as many primate species as possible. I received their final set of sequences in March 2013. I then performed the analyses presented below.

I presented part of these results with a poster at the Jacques Monod Conference "Recent advances on the evolution of sex and genetic systems", in Roscoff in May 2013.

A paper is in preparation. We plan to complete the work mentionned in the Perspectives below, before submitting a manuscript.

Evidence for frequent X-Y gene conversion events in primates in three evolutionary strata

Eugénie Pessia[1] · **Laurent Guéguen**[1] · **Victor Colomina**[2] · **Brigitte Crouau-Roy**[2] · **Emilie Lecompte**[2,*] · **Gabriel AB Marais**[1,*]

1 Laboratoire de Biométrie et Biologie Évolutive, Université Lyon 1, Centre National de la Recherche Scientifique, Villeurbanne F-69622 cedex, France; **2** Université de Toulouse, CNRS, UPS, EDB (Laboratoire Evolution et Diversité Biologique), 118 route de Narbonne, F-31062 Toulouse, France

Abstract The mammalian Y chromosome has long been considered a no recombination's land, this absence of recombination being responsible for the observed profound Y degeneration. Challenging this view, reports of both Y-Y and X-Y gene conversion have been made during the last decade. While Y-Y gene conversion has been studied in details in human, chimpanzee and macaque, studies of X-Y gene conversion have included few species and/or few genes. Here we present a detailed study of X-Y gene conversion between five gametologs belonging to the three most recent evolutionary strata in many primate species. Unexpectedly, we found at least one gene conversion event in each of these genes. In three of the genes (*Amel*, *Nlgn4*, *Prk*), the small detected regions undoubtedly underwent multiple gene conversion events during primate evolution. We discuss several interpretations of this phenomenon: a beneficial role of X-Y gene conversion for the maintenance of Y-linked genes, or the possibility that this mechanism has been conserved neutrally.

Introduction

The mammalian X and Y sex chromosomes are very different between each other in terms of length and structure (see e.g. [1]). Yet, in the ancestor of all mammals 210 million years ago (mya) they were a pair of autosomes [2,3], thus being nearly identical. When the therians (marsupials and placentals) ancestor acquired the male-determining gene *Sry* ∼180 mya, it is hypothesized that selection for recombination arrest in the *Sry* area took place in order to prevent genetic exchange by crossing overs (COs) between these proto-X and proto-Y chromosomes [4]. In the area where no genetic exchange happened anymore between both chromosomes, they started to diverge from each other and the Y-specific regions underwent degeneration by Hill-Robertson effects because of the absence of recombination (reviewed in [5]). This first recombination arrest was followed by others, gradually enlarging the sex-specific parts of both chromosomes and thus shortening the "pseudo-autosomal regions" (PARs). In humans, five different "evolutionary strata" have been identified, the oldest being stratum 1 close to PAR2 (Xq), and the youngest stratum 5 close to PAR1 (Xp) [6,7] (but see [8,9] for possible additionnal strata in human). In the non-recombining part of the sex chromosomes, genes that were formely allelic are called "gametologs", their sequence evolution being now supposedly independent.

The view that the Y-specific part of the Y chromosome is non-recombining has been challenged by the discovery of ampliconic regions (highly repeated regions) on the human, chimpanzee and macaque Y chromosomes, where ongoing Y-Y gene conversion occurs at elevated rates [10,11,12]. Gene conversion is a type of recombination happening in both CO and non-CO events where a non-reciprocal transfer of genetic information occurs, i.e. a "donor" sequence copies itself on a homologous "receptor" sequence after this latter underwent a double-strand break (see e.g. [13]). The formation of these amplicons has been hypothesized to be selected for, as Y-Y gene conversion between copies is a form of recombination thus slowing down Muller's ratchet and improving selection efficacy [10]. Theoretical studies indeed showed that it would be the case under high rates of Y-Y gene conversion [14,15].

Another type of gene conversion, where one X-linked gene recombines with its Y gametolog, has been reported several times in the last decade or so [20,16,21, 17,19,22,23,24]. One of the first genes in which X-Y gene conversion was found is *Amel* [25]. These authors attributed their results as the footprint of an ancient

* Corresponding authors: emilie.lecompte@univ-tlse3.fr · gabriel.marais@univ-lyon1.fr

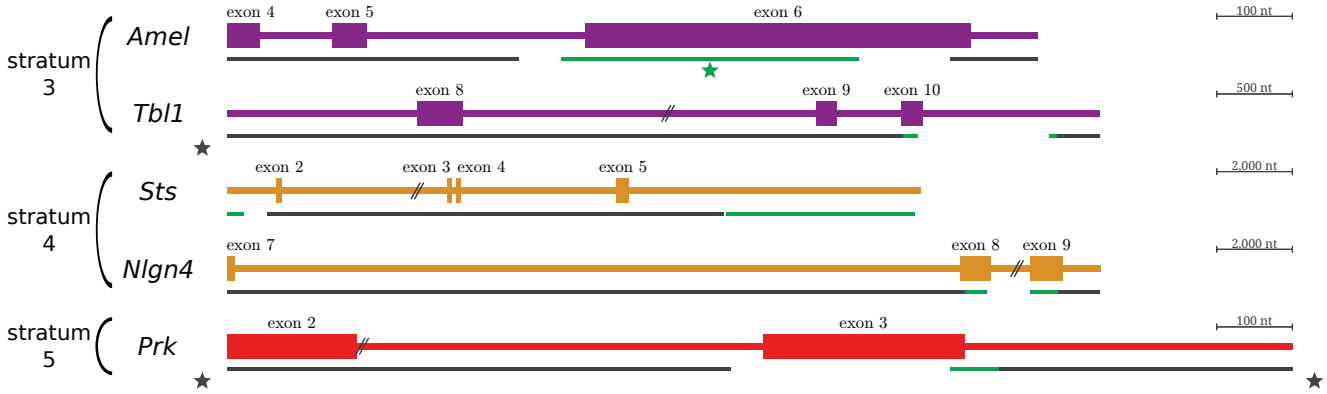


Fig. 1 Regions with gene conversion in five gametologs encompassing three evolutionary strata. Representation of the alignments of the five gametologs studied: *Amel* and *Tbl1* in human stratum 3 (purple), *Sts* and *Nlgn4* in stratum 4 (yellow), and *Prk* in stratum 5 (red). The positions of the exons and introns, and of the segments that probably underwent gene conversion (green) as well as the segments that probably did not undergo gene conversion (black) are shown to scale. Note that these positions do not correspond to real positions in any species as they are the positions in the alignments (available in Supplementary Materials). Stars were put when evidence for X-Y gene conversion was already known from previous work: in the same segment as us (green star [16]) or elsewhere in the gene (black stars [17,18,19]) The '//' symbol highlights parts of introns that were not sequenced in this study.

pseudo-autosomal boundary inside this gene [25]. But it was later shown that the 3' region of *Amel* underwent gene conversion while the 5' region did not [16], although *Amel* was probably close to the ancient boundary between the PAR and stratum 3 [22]. An adaptive role for X-Y gene conversion events was hypothesized by some authors [20,17,19], as the X-linked genes could help slowing down the degradation of their Y gametologs by X-to-Y gene conversion. However, no theoretical study has been done to support this idea.

Evidence for ancient X-Y gene conversion events have been found by exploring the phylogeny of small regions having detectably low X-Y divergence (using the "p-distance") at the scale of eutherians [16], primates [23] and felidae [20]. Indeed, if one pair of gametologs is located on a stratum created before the divergence between species A and B, then we expect the phylogeny to group separately X and Y copies from both species, i.e. the tree will be of the form: (A_X,B_X),(A_Y,B_Y). But for X-Y gene converted regions, species-specific gene conversion happened at some point in evolution and erased past divergence between X and Y sequences. The present-time tree obtained is thus: (A_X,A_Y),(B_X,B_Y). Other studies used SNP data and found signatures of very recent gene conversion events in human populations [17,19,24]. In all these studies, only few gene conversion events were analyzed, in a limited number of species (except in [23], but they had only two representatives of New-World Monkeys and one Old-World Monkey). Our goal here was to search for X-Y gene conversion regions at different loci along the X chromosome. For this, we sequenced five gametologs (*Amel*, *Tbl1*, *Sts*, *Nlgn4*, *Prk*) spanning the three most recent evolutionary strata, in a large number of primate species. We found strong evidence for X-Y gene conversion events in all of these five genes, several of these having undergone both recent and ancient gene conversion events, as well as events occuring in both X-to-Y and Y-to-X directions. Our results strongly support the view that the so-called non-recombining region of the Y chromosome (NRY) has actually recombined regularly during its evolution, by non-CO non-reciprocal events. We discuss the possibility that these gene conversion events were not selected for to slow down the degeneration of essential Y-linked genes: they might have occurred in a neutral manner during Y chromosome evolution.

Results

In order to study gene conversion between primates X and Y chromosomes along time, we sequenced five genes located on different evolutionary strata. We focused on strata 3 to 5, as the first two strata are very ancient, they contain few gametologs and all of them have dS values >1 (both stratum 1 and stratum 2 are shared between all therians (∼180 myo) [6,3]). We thus sequenced *Amel* and *Tbl1* on stratum 3 which was formed before the eutherian radiation (∼100 myo) [6], *Sts* (pseudogenized in human Y) and *Nlgn4* on stratum 4 formed in Simians (∼40 myo, although some authors exclude New-World Monkeys [1]), and *Prk* on stratum 5 recently formed in Catarrhini (∼30 myo) [7,12] (Figure 1). These five genes were chosen based on the criteria that they are not part of human, chimpanzee nor macaque amplicons [21,11,12]. When available we used sequences from public databases, and we obtained the

rest of the sequences by PCR amplification (see Materials and Methods). Briefly, we designed our primers based on the known sequences from sequenced primates (fully assembled chromosomes, contigs or BACs). Thus, for species belonging to clades where few sequences are publicly available we were less likely to amplify something. We were not able to amplify any Y-linked sequence and only few X-linked sequences in Prosimians, as expected. Strata 4 and 5 identified in human are indeed still part of the PAR in Prosimians (lemurs at least, see [26]) so only one sequence is expected for those genes, and for stratum 3 few Prosimian BACs were available for primer designing. We thus focus on Simian X-Y gene conversion. The characteristics of the sequences obtained are listed in Table S1. The alignments obtained are shown in Figure 1. Fewer Y than X chromosome primers amplified, as no Y chromosome is sequenced except for three very closely related species: human, chimpanzee and macaque [21,11,12]. In all five genes, we were able to obtain enough X and Y sequences to look for gene conversion events.

Using `Geneconv` [27], we were able to find signals of X-Y gene conversion in all the gametologous pairs studied (Figure 1) (see Materials and Methods). These results, together with the existing litterature, show that X-Y gene conversion is not an unfrequent phenomenon and has had a strong influence on X- and Y-linked gene evolution by wiping out their divergence accumulated since their evolutionary stratum was formed.

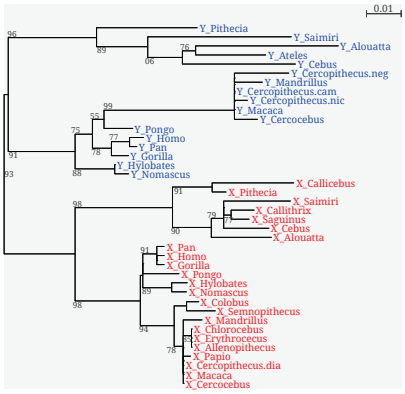
In order to further analyse the events that happened at each identified site, we computed the ML trees of the combined regions showing no footprint of gene conversion (NCR) and of the regions where gene conversion was detected (CR) (Figure 2). The trees of the NCRs follow remarkably well the species tree for the X and Y sequences separately, given the relatively small size of the regions under study. It confirms that these five gametologs do not have an evolutionary history different from what is expected in their respective evolutionary strata. The trees of the CRs, on the contrary and as expected, do not follow the normal phylogeny of their stratum. The different events deduced, and their direction when possible to determine, are summarized in Figure 3. The bootstrap values of the branches supporting these events are remarkably high, again given the small size of the studied regions (length without gaps in human: *Amel* 315 nt, *Tbl1* 98 and 53, *Sts* 371 and 3600, *Nlgn4* 554 and 708, *Prk* 65). The number of events inferred is large, and it is probably an underestimation. First, we do not know if only one event or multiple events are responsible for each dot on Figure 3. Second, human gene conversion tracts are typically 200 bp to 1 kb long [28], and one of the CRs we detected, the second CR in *Sts*, is notably longer than that (4922 nt). Thus, probably more events than the two we inferred happened in this region. Third, for some genes (*Tbl1*, *Sts*) the number of species representing each primate clade is small, thus when we have inferred an event at one branch it might represent multiple events of gene conversion. Moreover, the fewer representatives of a primate clade we have, the fewer number of events will be possible to infer in this clade. These two genes are indeed the ones were the least events could be inferred. For *Amel*, *Nlgn4* and *Prk*, on the contrary, X-Y gene conversion events are widely spread along the primate phylogeny, showing that they underwent multiple events of gene conversion.

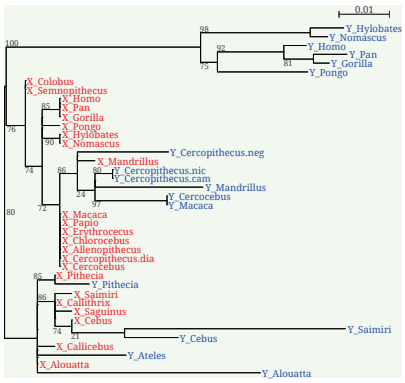
Discussion

Several previous studies on X-Y gene conversion concluded that this type of recombination may have been adaptive [20,17,19] and may have promoted the "repair" of essential Y-linked genes by their X-linked gametolog. However, our results do not strongly support this hypothesis, because our gene conversion boundaries are remarkably clear-cut and the other segments follow perfectly the phylogeny expected with an absence of X-Y gene conversion. It appears difficult to explain the presence of gene conversion in some exons only while conservation of the whole gene is probably needed. But stronger constrains on some exons may explain this pattern. Moreover, two events we detected might be in contradiction with this theory: 1) one gene conversion was from the Y to the X chromosome (in *Nlgn4*, Figure 3), however the direction of our detected gene conversion events is strongly biased toward X-to-Y; 2) one CR is entirely located in an intron (in *Sts*, Figure 1), but there may be a functional element strongly selected in this intron. Finally, theoretical papers showed that Y-Y gene conversion can interfere with degeneration only at elevated rates [14,15]. X-Y gene conversion rates are probably not this high as otherwise the phylogenies of the CRs would cluster the X and Y gametologs of every species. Indeed, estimated Y-Y gene conversion rates are $\sim 2.2 \times 10^{-4}$ per base per generation [10] and the average base mutation rate on the Y is $\sim 2.3 \times 10^{-8}$ [31]. The available estimates of X-Y gene conversion rates are lower than the ones of Y-Y events (lower bound: 3.8×10^{-8}, upper bound: 8.2×10^{-6} [18, 32]).

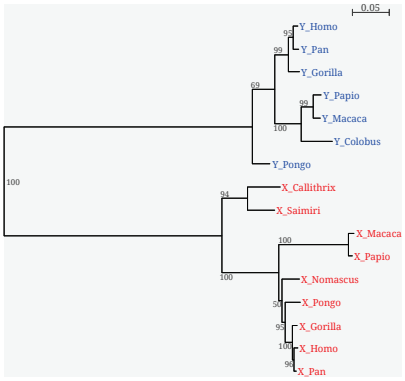
We hypothesize that there is at least one region of frequent X-Y gene conversion events in almost all gametologs simply by chance. In human autosomes, there might be one gene conversion hotspot every ∼600 bp [33,34]. Recent papers have shown that recombination

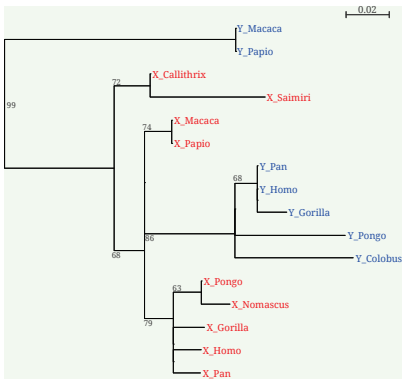
Amel

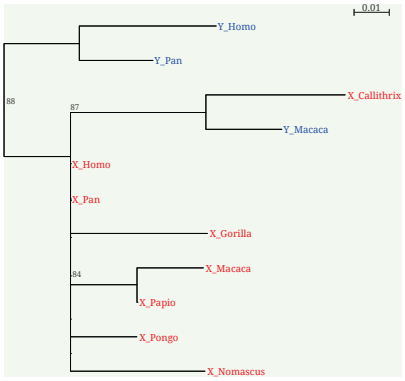




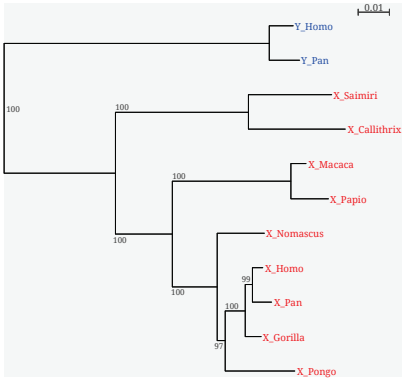
Tbl1

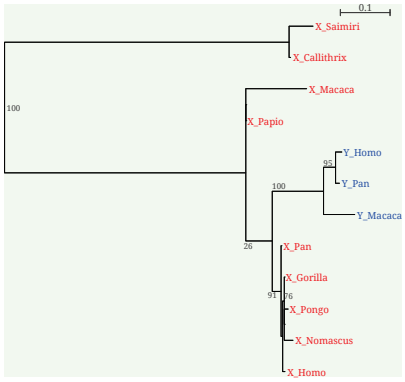


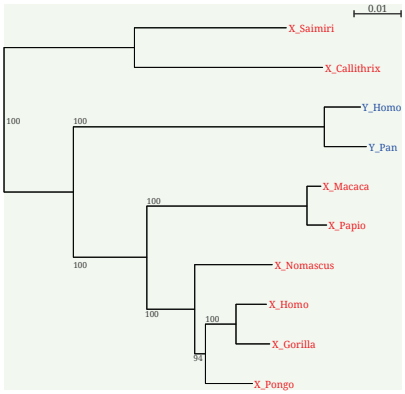




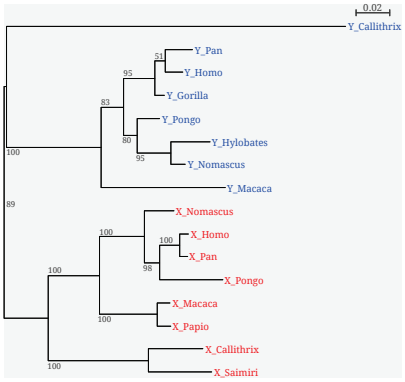
Sts

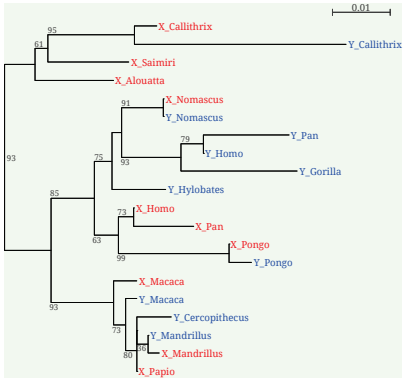


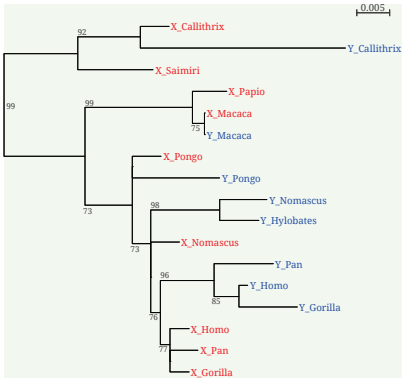




Nlgn4







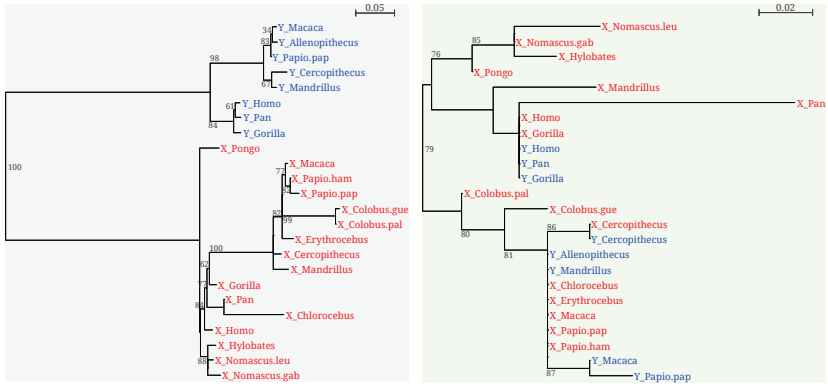


Fig. 2 Phylogenetic trees of the NCR and CR. Maximum-likelihood trees obtained for the NCRs (regions where no X-Y gene conversion was detected, black background) and the CRs (regions where gene conversion was detected, green background) are shown for *Amel*, *Tbl1*, *Sts*, *Nlgn4* and *Prk* (see Figure 1 and Materials and Methods). The X-linked sequences are highlithed in red, and the Y-linked sequences in blue. Bootstrap values (in %) are indicated when they were different from zero. Branch length scale (number of substitutions per site) is indicated on each tree. Due to gaps in the sequences, the species set can be different between trees of a same gene (see Materials and Methods).

mechanisms are highly different in the PAR and in autosomes in mice (reviewed in [35]). But we can hypothesize that before the formation of a stratum when the X and Y regions are still pseudo-autosomal, X-Y gene conversion probably occurs at several hotspots all along this region. After the stratum formation, X and Y regions start to diverge. In genes, that are under purifying selection (despite this selection being lowered by the absence of recombination [5]), these gene conversion hotspots might be maintained neutrally, allowing for X and Y pairing after a double-strand break (DSB). Alternatively, if there was no prior gene conversion hotspot, a gene conversion event might have happened by chance at a random location when the X and Y sequences where not too much diverged to pair, then after a time this region will be the only one prone to future gene conversion events, the rest of the gene being too diverged (usually >95% homology between the interacting sequences of a gene conversion event [28]). The location of this first gene conversion event could be non-random, and favored by an element such as transposable elements (TEs) or PRDM9 sequence motif [36] (but see [37] showing COs are independent of PRDM9 in the mouse PAR). Iwase and collaborators indeed found a LINE insertion at the proximal end of their CR in KAL and suspected it to be responsible for the gene conversion events [23]. However, it has been shown that Alu and L1 TEs can be inserted in pre-existing DSBs [38], meaning that the LINE insertion found by [23] may have been the consequence and not the cause of these gene conversion events.

The fact that more X-to-Y than Y-to-X events were observed, in this study and others, could be seen as an argument in favor of X-Y gene conversion being adaptive. However, several types of biases could lead to an apparent excess of X-to-Y gene conversion events, when there is in fact no selection for a "repair" of Y copies by their X homologs. First, our sequence data is biased toward X sequences, because of several technical difficulties in designing Y-specific primers (see Materials and Methods). Second, the Y chromosome may be more prone to DSBs than the X, thus favoring X-to-Y gene conversion. 1) This could be due to a fragility of the Y chromosome during replication, for example because of its high heterochromatinization or of its high TE density. 2) DSBs can be caused by transposon insertions (P-element in *Drosophila* [39]), thus leading to gene conversion events. The same amount of TE insertions occurs between the X and the Y, but because of reduced selection on the Y [5] insertions on this chromosome are less often removed from the population than the ones on the X. This means that TE insertion events on the X, possibly leading to a Y-to-X gene conversion, will be often removed from the population while the events possibly leading to X-to-Y gene conversion will not. This may explain why we see almost only X-to-Y gene conversion events. Third, hypothesizing that there is no selection for X-to-Y gene conversion does not imply that there is no counter-selection for Y-to-X events. As selection is more efficient on the X chromosome, whenever a Y-to-X gene conversion that copies a deleterious mutation occurs it will be selectively removed from the population, thus creating a bias in the fixed events of gene conversion.

If X-Y gene conversion can occur in gametologs, we expect the genes present in the same stratum, despite having stopped recombining (by COs) at the same time, to present a variation in their X-Y dS. This could explain why human strata 4 and 5 have very different dS despite their estimated age differing by only 10 my

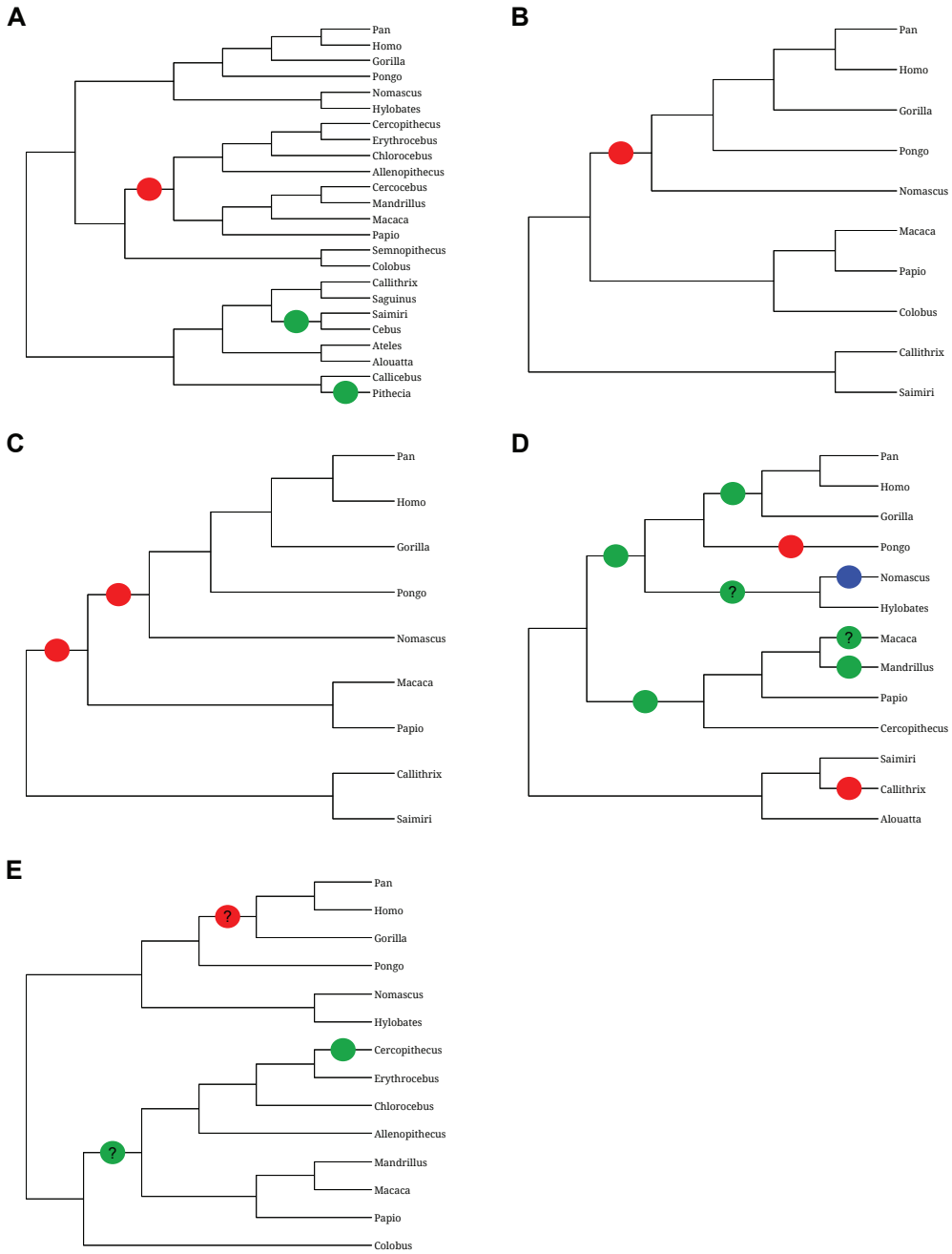


Fig. 3 Predictions of the minimal number of gene conversion events. The gene conversion events between X and Y chromosomes, inferred from Figure 2, are showed for each gene: *Amel* (**A**), *Tbl1* (**B**), *Sts* (**C**), *Nlgn4* (**D**) and *Prk* (**E**). Red dots are used for events where the X chromosome sequence copied itself on the Y (X-to-Y), blue dots symbolize the opposite type of events (Y-to-X), and green dots were put when a direction could not be assigned with certainty. Question marks point out cases where the branches in the Figure 2 trees that led to hypothesize a gene conversion event are supported by low bootstrap values. The reference primate phylogeny used is from both [29,30].

(stratum 4 is 40 myo, stratum 5 is 30 myo). Stratum 4 gametologs are indeed two-times more diverged than stratum 5 ones (∼10% vs ∼5%) [7]. Hughes and collaborators hypothesized a difference in the mutation rates between these two strata to explain this two-fold dS difference [12], but we suggest that they are rather due to more recent gene conversion events in the last stratum, where some intergenic gene conversion may still be occurring, than in older strata. In the unique gene in stratum 5 that we analyzed, *Prk*, less gene conversion events were inferred than in e.g. *Nlgn4* (three vs nine,

Figure 3). This may be an underestimation, as the tree of the CR presents multiple polytomies (see Figure 2).

Perspectives

Future studies on X-Y gene conversion will consist in the caracterisation of the impact of X-Y gene conversion at the molecular evolution level. Gene conversion is actually biased in mammals ([40] and references therein). The BGC (Biased Gene Conversion) locally increases the GC* (GC% that a sequence will reach at equilibrium) [41,42]. We thus expect a higher GC* in the CRs compared to other regions of gametologs. Using maximum-likelihood methods such as in the BppML software [43], it would thus be interesting to compute the GC* values of CRs and NCRs in order to estimate the rate of X-Y gene conversion events. A study found no footprint of BGC on the human sex chromosomes (outside PARs) compared to autosomes [44], but these authors used a large-scale approach which probably did not allow to detect very localized gene-converted regions. They did find an elevated rate of X-Y "biased clustered substitutions" close to PAR1 [44], i.e. in the most recent stratum. Second, one could study the effects of X-Y gene conversion on the intensity of selection affecting these gametologs. Indeed, if the GC% increases because of BGC, dN/dS values will increase as well because non-synonymous AT→GC mutations can be fixed, thus mimicking a positive selection process even if those mutations are deleterious [45]. This means that X-Y gene conversion could not only have no selected role in the "repair" of Y-linked genes, but could also increase Y-linked genes degeneration by fixing deleterious AT→GC mutations. However, the number of gene conversion events during evolution might be too low for this effect to be important. But theoretical work is needed in order to study the effect of X-Y gene conversion on Y evolution.

Materials and Methods

DNA sampling and sequences alignment

DNA materials have very variable origins: 1) from public databases, for species where the X-linked and/or Y-linked copies have been sequenced; 2) from faeces of individuals in the wild; 3) from hairs, blood or biopsies (ear) of individuals from zoos.

For the sequences that were not available in public databases, primers were designed for PCR amplifications. Primer design was based on the available data in other species: either fully assembled chromosomes, contigs or BACs. For some sequences, especially Y-linked genes, a great number of primers had to tested before being able to amplify any material, and sometimes amplification was never obtained. This is due to several reasons: 1) Very few Y chromosomes are fully sequenced and these are very closely related to each other (human, chimpanzee and macaque [21,11,12]). Moreover, due to both male-biased mutation rates [46] and weakened selection [5], the Y chromosome evolve faster than the X and autosomes. Thus, designing primers was difficult for Y-linked copies outside of Apes. 2) Because of the bias in the available data towards X-linked sequences, in non-Ape species it was difficult to design primers specific to Y-linked sequences. We sometimes amplified the X-linked copy of the gametologous pair. 3) Four of our five genes are not pseudogenized in human (Y-linked *Stsp1* is a human pseudogene), but they may be pseudogenized in some primates and thus impossible to amplify.

The sequences obtained for each gene were aligned using `Sequencer` with the `Muscle` software [47], and these alignments were double-checked visually.

Conversion tracts detection and trees construction

The nucleotidic sites that possibly underwent gene conversion were determined using `Geneconv` v1.81 [27]. The species for which much fewer nucleotides had been sequenced were removed from the alignments, as `Geneconv` removes from its analysis every site in which at least one sequence has a gap. The *gscale* argument was set to 2 in order to allow a reasonable amount of mismatches in the alignments. The nucleotidic model was used for the whole alignments (exons plus introns). `Geneconv` was run using the *group* option, specifying each species for which both an X and a Y sequences were available: these species were the ones in which `Geneconv` had to look for gene conversion. Whenever `Geneconv` found a segment of gene conversion in one pair of X and Y sequences, this segment was considered as a possible gene conversion region (CR) for all the species of the alignment. When several segments were found for a same gene, either in the same species or in different pairs of X-Y sequences, we considered them as independent CRs if their borders did not overlap and were distant, otherwise we took the smallest common segment as a probable CR and considered the rest of the sites detected by `Geneconv` as equivocal. All the nucleotides that are not in a CR nor in an equivocal segment are put together to form a region showing no conversion (NCR). See Figure 1 for the CRs and NCRs boundaries obtained for each of the five genes.

The phylogenetic trees shown in Figure 2 were computed using PhyML in `Seaview` with default parameters (model: GTR, invariable sites: none, across rate variation: optimized, tree searching operations: NNI, starting tree: BioNJ optimizing tree topology), on CRs and NCRs separately, as defined above. Boostraping was performed in `Seaview` using an aLTR method. The NCR trees were made using all the NCRs of one gene combined together. The CR trees, on the contrary, were computed for each CR separately. For each tree, only the sequences containing few enough gaps in the concerned region to construct the ML-tree were used, thus the used sequences can differ between trees of a same gene.

Acknowledgements This work was supported by the ANR grant "SexPrim" (ANR-12-BSV7-0002).

References

1. Hughes JF, Rozen S (2012) Genomics and genetics of human and primate y chromosomes. Annu Rev Genomics Hum Genet 13: 83–108.
2. Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, et al. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 6: e80.
3. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res 18: 965–973.
4. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb) 95: 118–128.
5. Charlesworth B, Charlesworth D (2000) The degeneration of y chromosomes. Philos Trans R Soc Lond B Biol Sci 355: 1563–1572.
6. Lahn BT, Page DC (1999) Four evolutionary strata on the human x chromosome. Science 286: 964–967.
7. Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al. (2005) The dna sequence of the human x chromosome. Nature 434: 325–337.
8. Pandey RS, Wilson Sayres MA, Azad RK (2013) Detecting evolutionary strata on the human x chromosome in the absence of gametologous y-linked sequences. Genome Biol Evol 5: 1863–1871.
9. Wilson MA, Makova KD (2009) Evolution and survival on eutherian sex chromosomes. PLoS Genet 5: e1000568.
10. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, et al. (2003) Abundant gene conversion between arms of palindromes in human and ape y chromosomes. Nature 423: 873–876.
11. Hughes JF, Skaletsky H, Pyntikova T, Graves TA, van Daalen SKM, et al. (2010) Chimpanzee and human y chromosomes are remarkably divergent in structure and gene content. Nature 463: 536–539.
12. Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Graves T, et al. (2012) Strict evolutionary conservation followed rapid gene loss on human and rhesus y chromosomes. Nature 483: 82–86.
13. de Massy B (2003) Distribution of meiotic recombination sites. Trends Genet 19: 514–522.
14. Connallon T, Clark AG (2010) Gene duplication, gene conversion and the evolution of the y chromosome. Genetics 186: 277–286.
15. Marais GAB, Campos PRA, Gordo I (2010) Can intra-y gene conversion oppose the degeneration of the human y chromosome? a simulation study. Genome Biol Evol 2: 347–357.
16. Marais G, Galtier N (2003) Sex chromosomes: how x-y recombination stops. Curr Biol 13: R641–R643.
17. Rosser ZH, Balaresque P, Jobling MA (2009) Gene conversion between the x chromosome and the male-specific region of the y chromosome at a translocation hotspot. Am J Hum Genet 85: 130–134.
18. Cruciani F, Trombetta B, Macaulay V, Scozzari R (2010) About the x-to-y gene conversion rate. Am J Hum Genet 86: 495–7; author reply 497-8.
19. Trombetta B, Cruciani F, Underhill PA, Sellitto D, Scozzari R (2010) Footprints of x-to-y gene conversion in recent human evolution. Mol Biol Evol 27: 714–725.
20. Pecon Slattery J, Sanner-Wachter L, O'Brien SJ (2000) Novel gene conversion between x-y homologues located in the nonrecombining region of the y chromosome in felidae (mammalia). Proc Natl Acad Sci U S A 97: 5307–5312.
21. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.
22. Lemaitre C, Braga MD, Gautier C, Sagot MF, Tannier E, et al. (2009) Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. Genome biology and evolution 1: 56.
23. Iwase M, Satta Y, Hirai H, Hirai Y, Takahata N (2010) Frequent gene conversion events between the x and y homologous chromosomal regions in primates. BMC Evol Biol 10: 225.
24. Sarbajna S, Denniff M, Jeffreys AJ, Neumann R, Soler Artigas M, et al. (2012) A major recombination hotspot in the xqyq pseudoautosomal region gives new insight into processing of human gene conversion events. Hum Mol Genet 21: 2029–2038.
25. Iwase M, Satta Y, Hirai Y, Hirai H, Imai H, et al. (2003) The amelogenin loci span an ancient pseudoautosomal boundary in diverse mammalian species. Proc Natl Acad Sci U S A 100: 5258–5263.
26. Gläser B, Myrtek D, Rumpler Y, Schiebel K, Hauwy M, et al. (1999) Transposition of sry into the ancestral pseudoautosomal region creates a new pseudoautosomal boundary in a progenitor of simian primates. Hum Mol Genet 8: 2071–2078.
27. Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6: 526–538.
28. Chen JM, Cooper DN, Chuzhanova N, Férec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8: 762–775.
29. Finstermeier K, Zinner D, Brameier M, Meyer M, Kreuz E, et al. (2013) A mitogenomic phylogeny of living primates. PLoS One 8: e69504.
30. Ranwez V, Berry V, Criscuolo A, Fabre PH, Guillemot S, et al. (2007) Physic: a veto supertree method with desirable properties. Syst Biol 56: 798–817.
31. Repping S, van Daalen SKM, Brown LG, Korver CM, Lange J, et al. (2006) High mutation rates have driven extensive structural polymorphism among human y chromosomes. Nat Genet 38: 463–467.

32. Rosser ZH, Balaresque PL, Jobling MA (2010) Response to cruciani et al., about the x-to-y gene conversion rate. Am J Hum Genet 86: 497–498.
33. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P (2005) A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.
34. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454: 479–485.
35. Kauppi L, Jasin M, Keeney S (2012) The tricky path to recombining x and y chromosomes in meiosis. Ann N Y Acad Sci 1267: 18–23.
36. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, et al. (2010) Prdm9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836–840.
37. Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV (2012) Genetic recombination is directed away from functional genomic elements in mice. Nature 485: 642–645.
38. Srikanta D, Sen SK, Conlin EM, Batzer MA (2009) Internal priming: an opportunistic pathway for l1 and alu retrotransposition in hominins. Gene 448: 233–241.
39. Engels WR, Johnson-Schlitz DM, Eggleston WB, Sved J (1990) High-frequency p element loss in drosophila is homolog dependent. Cell 62: 515–525.
40. Romiguier J, Ranwez V, Douzery EJP, Galtier N (2010) Contrasting gc-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res 20: 1001–1009.
41. Duret L, Galtier N (2009) Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10: 285–311.
42. Lesecque Y, Mouchiroud D, Duret L (2013) Gc-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. Mol Biol Evol 30: 1409–1419.
43. Dutheil J, Boussau B (2008) Non-homogeneous models of sequence evolution in the bio++ suite of libraries and programs. BMC Evol Biol 8: 255.
44. Dreszer TR, Wall GD, Haussler D, Pollard KS (2007) Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res 17: 1420–1430.
45. Galtier N, Duret L, Glémin S, Ranwez V (2009) Gc-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet 25: 1–5.
46. Wilson MA, Makova KD (2009) Genomic analyses of sex chromosome evolution. Annu Rev Genomics Hum Genet 10: 333–354.
47. Edgar RC (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792–1797.

Supporting Information

Table S1 List of primates sequenced for each gene

Gene name	Evol. strata[a]	Chromosome	Species	Regions sequenced[b]	Length
AMEL	3	X	Homo sapiens	exon4 to intron6	959
		X	Pan troglodytes	exon4 to intron6	959
		X	Gorilla gorilla	exon4 to intron6	959
		X	Pongo pygmaeus	exon4 to intron6	959
		X	Hylobates lar	exon5 to exon6	721
		X	Nomascus leucogenys	exon4 to intron6	959
		X	Colobus guereza	exon4 to intron6	928
		X	Semnopithecus entellus	exon5 to exon6	736
		X	Allenopithecus nigroviridus	exon4 to intron6	904
		X	Erythrocebus patas	exon4 to intron6	905
		X	Cercopithecus diana	intron5 to exon6	641
		X	Chlorocebus tantalus	exon4 to intron6	909
		X	Papio papio	exon4 to intron6	943
		X	Mandrillus sphinx	exon4 to intron6	861
		X	Cercocebus atys	exon4 to intron6	918
		X	Macaca mulatta	exon4 to intron6	959
		X	Pithecia pithecia	intron4 to exon6	737
		X	Callicebus cupreus	exon4 to intron6	901
		X	Alouatta macconnelli	intron4 to intron6	843
		X	Saguinus midas	intron4 to intron6	878
		X	Callithrix geoffroyi	exon4 to intron6	964
		X	Cebus apella	intron4 to intron6	879
		X	Saimiri sciureus	exon4 to intron6	963

Table S1 (continued)

Gene name	Evol. strata[a]	Chromosome	Species	Regions sequenced[b]	Length
		Y	Homo sapiens	exon4 to intron6	961
		Y	Pan troglodytes	exon4 to intron6	961
		Y	Gorilla gorilla	intron5 to exon6	693
		Y	Pongo pygmaeus	exon5 to intron6	745
		Y	Hylobates lar	exon5 to intron6	738
		Y	Nomascus gabriellae	exon5 to intron6	734
		Y	Cercopithecus campbelli	intron5 to exon6	631
		Y	Cercopithecus neglectus	exon5 to exon6	679
		Y	Cercopithecus nictitans	exon5 to intron6	723
		Y	Mandrillus sphinx	exon5 to intron6	730
		Y	Cercocebus atys	intron5 to exon6	642
		Y	Macaca mulatta	exon4 to intron6	952
		Y	Pithecia pithecia	intron4 to exon6	779
		Y	Alouatta macconnelli	intron4 to intron6	839
		Y	Ateles paniscus	intron4 to intron6	900
		Y	Cebus apella	intron4 to exon6	782
		Y	Saimiri sciureus	intron4 to exon6	728
TBL1	3	X	Homo sapiens	intron7 to intron10	4,256
		X	Pan troglodytes	intron7 to intron8 and exon10 to intron10	3,438
		X	Gorilla gorilla	intron7 to intron10	3,995
		X	Pongo pygmaeus	intron7 to intron10	4,219
		X	Nomascus gabriellae	intron7 to intron10	3,923
		X	Papio papio	intron7 to intron10	4,321
		X	Macaca mulatta	intron7 to intron10	4,140
		X	Callithrix geoffroyi	intron7 to intron10	3,272
		X	Saimiri sciureus	intron7 to intron10	3,392
		Y	Homo sapiens	intron7 to intron10	3,920
		Y	Pan troglodytes	intron7 to intron10	4,062
		Y	Gorilla gorilla	intron7 to intron10	1,351
		Y	Pongo pygmaeus	intron7 to intron10	1,339
		Y	Colobus guereza	intron7 to intron10	1,218
		Y	Papio papio	intron8 to intron10	725
		Y	Macaca mulatta	intron7 to intron10	2,673

Table S1 (continued)

Gene name	Evol. strata[a]	Chromosome	Species	Regions sequenced[b]	Length
STS	4	X	Homo sapiens	intron1 to intron5	11,778
		X	Pan troglodytes	intron1 to intron5	11,578
		X	Gorilla gorilla	intron1 to intron5	11,665
		X	Pongo pygmaeus	intron1 to intron5	12,513
		X	Hylobates lar	intron2 to exon5	993
		X	Nomascus leucogenys	intron1 to intron5	11,785
		X	Nomascus gabriellae	intron2 to exon5	1,215
		X	Colobus guereza	intron2 to exon4	659
		X	Presbytis cristatus	intron2 to exon4	659
		X	Cercopithecus ascanius	intron2 to exon4	659
		X	Cercopithecus diana	intron2 to intron5	1,212
		X	Papio papio	intron1 to intron5	10,654
		X	Mandrillus sphinx	intron2 to exon5	998
		X	Macaca mulatta	intron1 to intron5	10,615
		X	Pithecia pithecia	intron4 to intron5	629
		X	Lagothrix sp	intron4 to exon5	560
		X	Callithrix geoffroyi	intron1 to intron5	11,501
		X	Saimiri sciureus	intron1 to intron5	11,156
		Y	Homo sapiens	intron1 to intron5	9,436
		Y	Pan troglodytes	intron1 to intron5	8,917
		Y	Hylobates lar	intron4 to exon5	325
		Y	Nomascus leucogenys	intron2 to exon4	624
		Y	Macaca mulatta	intron1 to intron2	1,907
NLGN4	4	X	Homo sapiens	exon7 to intron9	8,234
		X	Pan troglodytes	exon7 to intron9	8,118
		X	Gorilla gorilla	intron8 to intron9	1,864
		X	Pongo pygmaeus	exon7 to intron9	4,406
		X	Nomascus leucogenys	exon7 to intron9	8,397
		X	Papio papio	exon7 to intron9	7,760
		X	Mandrillus sphinx	intron7 to intron8	892
		X	Macaca mulatta	exon7 to intron9	8,276
		X	Alouatta macconnelli	intron7 to intron8	881
		X	Callithrix geoffroyi	exon7 to intron9	6,993
		X	Saimiri sciureus	exon7 to intron9	8,206

Table S1 (continued)

Gene name	Evol. strata[a]	Chromosome	Species	Regions sequenced[b]	Length
		Y	Homo sapiens	exon7 to intron9	8,369
		Y	Pan troglodytes	exon7 to intron9	7,626
		Y	Gorilla gorilla	exon7 and exon8 to intron9	2,107
		Y	Pongo pygmaeus	intron7 to intron9	2,014
		Y	Hylobates lar	intron7 to intron9	2,047
		Y	Nomascus gabriellae	intron7 to intron9	2,018
		Y	Cercopithecus diana	intron7 to intron8	933
		Y	Mandrillus sphinx	intron7 to intron8	913
		Y	Macaca mulatta	exon7 to intron9	8,615
		Y	Callithrix geoffroyi	exon7 to intron9	19,322
PRK	5	X	Homo sapiens	exon2 and intron2 to intron3	1,035
		X	Pan troglodytes	exon2 and intron2 to intron3	660
		X	Gorilla gorilla	exon2 and intron2 to intron3	1,035
		X	Pongo pygmaeus	exon2 and intron2 to intron3	1,165
		X	Hylobates lar	exon2 and intron2 to intron3	651
		X	Nomascus gabriellae	exon2 and intron2 to intron3	713
		X	Nomascus leucogenys	exon2 and intron2 to intron3	950
		X	Colobus angolensis	intron2 to intron3	573
		X	Colobus guereza	intron2 to intron3	573
		X	Allenopithecus nigroviridus	exon2	169
		X	Erythrocebus patas	intron2 to intron3	536
		X	Cercopithecus diana	intron2 to intron3	517
		X	Chlorocebus tantalus	intron2 to intron3	437
		X	Papio hamadryas	exon2 and intron2 to intron3	839
		X	Papio papio	intron2 to intron3	563
		X	Mandrillus sphinx	exon2 and intron2 to intron3	615
		X	Macaca mulatta	exon2 and intron2 to intron3	943

Table S1 (continued)

Gene name	Evol. strata[a]	Chromosome	Species	Regions sequenced[b]	Length
		Y	Homo sapiens	exon2 and intron2 to intron3	1,320
		Y	Pan troglodytes	exon2 and intron2 to intron3	1,321
		Y	Gorilla gorilla	exon2 and intron2 to intron3	1,001
		Y	Nomascus gabriellae	exon2	169
		Y	Allenopithecus nigroviridus	intron2 to intron3	644
		Y	Cercopithecus diana	intron2 to intron3	766
		Y	Papio papio	intron2 to intron3	628
		Y	Mandrillus sphinx	intron2 to intron3	710
		Y	Macaca mulatta	exon2 and intron2 to intron3	1,215

[a] as described in [7]

[b] exons and introns numbers are indicated using human genes as a reference. Entire introns have not been sequenced in this study, we restrained ourselves to the parts close to exons. For example, exon2 to intron3 means that we sequenced both the 3' and 5' ends of the intron number 2, and only the 5' part of intron 3.

3

Sex chromosomes of a third kind: UV system evolution in a billion-year-distant brown alga, *Ectocarpus siliculosus*

A team of molecular developmental biologists working on a brown alga, *Ectocarpus siliculosus*, contacted Gabriel Marais four years ago for his expertise in molecular evolution of sex chromosomes. This group from Roscoff (Britany) had already sequenced the genome of a haploid male individual of *Ectocarpus* [Cock et al. *Nature* 2010]. They wanted to study the sex chromosomes of a UV type in this species. Together with Gabriel and other teams, they initiated a common project where:

- the Genoscope (Évry, France) sequenced by NGS a female haploid genome and performed its assembly.
- Ghent (Belgium) annotated this genome.
- Roscoff determined the sex-specific region boundaries by comparing the male and female haploid genomes, and confirmed them by PCRs and genetic maps. They sequenced the transcriptome by NGS at different phases of the life cycle. They also produced triploid and tetraploid mutants by cell culture.
- Lyon made the molecular evolution analyses: TE accumulation, codon usage bias

analyses, estimation of the sex chromosome age by dS analyses.

Being interested in the evolution of sex chromosomes in non-model organisms, I asked to be part of this project. I made a significant contribution on all the molecular evolution analyses, allowing me to be one of the three co-first authors of the resulting paper.

I gave a selected talk on the molecular evolution results of this project at the SMBE annual meeting, at Chicago in July 2013.

After being submitted to several journals, this paper was submitted to *Current Biology* on the 11^{th} of October 2013 and is currently under review.

An ancient system of haploid sex determination in a distant eukaryote

Ahmed S[1,2,§] · **Cock JM** [1,§] · **Pessia E**[3,§] · **Luthringer R**[1] · **Cormier A**[1] · **Robuchon M**[1,8] · **Sterck L**[4] · **Peters AF**[5] · **Dittami SM**[1] · **Corre E**[7] · **Valero M**[8] · **Aury J-M**[6] · **Roze D**[8] · **Van de Peer Y**[4] · **Bothwell JH**[2] · **Marais GAB**[3] · **Coelho SM**[1,*]

1 UPMC Univ Paris 6 – CNRS UMR 7139 "Marine Plants and Biomolecules", Station Biologique de Roscoff, CS 90074 29688 Roscoff, France; **2** Queens University Belfast, Medical Biology Centre, Belfast BT9 7BL, Northern Ireland, UK; **3** Université Lyon 1, Centre National de la Recherche Scientifique, UMR 5558, Laboratoire de Biométrie et Biologie Évolutive, 69622 Villeurbanne, France; **4** Department of Plant Systems Biology, VIB, Technologiepark 927, B-9052 Gent, Belgium; **5** Bezhin Rosko, 29250 Santec, France; **6** Commissariat à l'Énergie Atomique (CEA), Institut de Génomique (IG), Genoscope, Evry, France; **7** Abims Platform, FR2424, Station Biologique de Roscoff, CS 90074, 29688 Roscoff, France; **8** UMR CNRS –UPMC 7144, Station Biologique de Roscoff, CS 90074, 29688 Roscoff, France

Abstract Background: A common feature of most genetic sex-determination systems studied so far is that they are regulated by either non-recombining chromosomal regions or by sex chromosomes, both of which have evolved independently and repeatedly across diverse species. A number of such sex-determining regions (SDR) have been studied in animals, plants and fungi, but very little is known about the evolution of sexes in other eukaryotic lineages.

Results: We report here the sequencing and genomic analysis of the sex chromosomes of *Ectocarpus*, a brown alga that has been evolving independently from plants, animals and fungi for over a billion years. In *Ectocarpus*, sex is expressed during the haploid phase of the life cycle, and both the female (U) and the male (V) sex chromosomes contain non-recombining regions. The U and V of this species have been diverging for more than 100 My, yet gene degeneration has been modest, the SDR has remained relatively small with no evidence for evolutionary strata. These features may be explained by the occurrence of strong purifying selection during the haploid phase of the life cycle and the low level of sexual dimorphism. V was dominant over U, suggesting that femaleness is the default state, adopted when the male haplotype is absent.

Conclusions: The *Ectocarpus* UV system has clearly had a distinct evolutionary trajectory not only to the well-studied XY and ZW systems, but also to the UV systems described so far. Nonetheless, some striking similarities exist, indicating remarkable universality of the underlying processes shaping sex chromosome evolution across distant lineages.

Highlights

- *Ectocarpus* U and V sex chromosomes evolved more than 100 MY ago
- The non-recombining region in the U and V is small and degeneration has been modest
- U and V are structurally similar but V is dominant over U
- Haploid selection and low sexual dimorphism may explain the sex chromosome structure

Introduction

Genetic determination of sex is mediated by extensive sex-determining regions (SDRs) or by sex chromosomes in a broad range of eukaryotes. Sex chromosomes have arisen independently and repeatedly across the eukaryotic tree and comparative analysis of different sex-determination systems has provided insights into how these systems originate and evolve. A typical sex chromosome pair is thought to derive from a pair of autosomes through the acquisition of genes involved in sex determination. If more than one locus involved in sex determination is located on the chromosome, recombination between loci is expected to be suppressed to avoid the production of mal-adapted individuals with a combination of male and female alleles of the sex-determining genes. This leads to the establishment of

§ Equal contribution
* Corresponding author: coelho@sb-roscoff.fr

a non-recombining region on the nascent sex chromosome, with important consequences for the evolution of this region of the genome [1]. For example, as a result of the suppression of recombination within the SDR repetitive junk DNA accumulates, leading to an increase in SDR size and degeneration of genes within the non-recombining region. At a later stage, deletion of non-functional DNA from within the SDR may lead to a decrease in the physical size of the SDR.

There is also evidence that the non-recombining region can progressively encroach on the flanking regions of the chromosome, so that it encompasses an increasingly greater proportion of the sex chromosome. This process is thought to be driven by the recruitment of genes with differential selective benefits to the two sexes (sexually antagonistic genes) into the SDR [2] (but see [3]). Extension of the SDR in this manner can lead to the creation of "strata" within the SDR corresponding to regions that have become non-recombining at different points in evolutionary time [4,5,6,7].

The genetic mechanism of sex determination also influences how the sex chromosomes evolve. In organisms where sex is expressed in the diploid phase, such as most animals and land plants, one sex is heterogametic (XY or ZW) whilst the other is homogametic (XX or ZZ). In these systems only the Y or W contain non-recombining regions because the X and Z recombine in the homogametic sex. In some algae and bryophytes the male and female sexes are genetically determined after meiosis, during the haploid phase of the life cycle [8]. This type of sexual system, termed UV to distinguish it from the XY and ZW systems described above [9], exhibits novel evolutionary and genetic properties that have no exact equivalent in diploid systems. In UV systems, the female and male SDR haplotypes function in independent, haploid individuals and consequently, there is no heterozygous sex comparable to XY males or ZW females. This difference between UV and XY/ZW systems should have important implications for SDR evolution [8]. In particular, neither the female U nor the male V SDRs recombine but degeneration of these regions is expected to be minimal provided that both contain genes that are essential during the haploid phase. Moreover, the U and the V SDR haplotypes should have similar characteristics because they function independently in two different individuals, and are therefore under similar evolutionary pressures [8]. Some asymmetry may be expected between the U and V, however, if sexual selection is stronger in males [10] or if one of the chromosomes plays a more active role in sex determination.

These verbal predictions of the characteristics of UV systems have been lacking empirical support. Although eukaryotic species with UV systems may be as common as those with XY and ZW systems, very few of the former have been characterised, with detailed sequence data being available for only two members of the Archeaplastidea lineage: the liverwort *Marchantia* (which has a fully sequenced V but an unidentified U chromosome) [11] and a UV pair of unknown age in the green alga *Volvox* [12], together with more fragmentary information recently obtained for the moss Ceratodon [13]. Clearly, additional detailed sequence information is required to fully test the predictions that have been made with respect to UV sex-determination systems and to evaluate the generality of these predictions in a broad phylogenetic context.

We report here the identification and the genetic and genomic characterisation of the U and V chromosomes of the brown algal model *Ectocarpus* [14,15]. Brown algae belong to the Stramenopiles, a lineage very distantly related to animals, fungi and green plants (the common ancestors dating back more than a billion years). The brown algae are considered to possess sex chromosomes rather than mating-type chromosomes [16,17,18] for a number of reasons: 1) there is a strict correlation between gamete size and sex in anisogamous species, 2) most sexual brown algal species exhibit some form of sexual dimorphism, [19,20] and 3) heteromorphic sex chromosomes have been identified [21,22]. Previous work has shown that sex is determined by a single, Mendelian locus in *Ectocarpus* [23]. During the haploid-diploid life cycle of this organism, meio-spores, produced by the sporophyte generation, develop into dioecious (separate male and female) gametophytes, which then produce either male or female gametes (Figure 1A).

We show here that the *Ectocarpus* UV has features typical of sex chromosomes in other systems such as low gene density and a large amount of repeated DNA. The male and female sex-determining regions (SDRs) are extremely diverged, reflecting a long independent evolutionary history, which we estimated at approximately 100-200 million years. Despite its age, the SDR has remained relatively small, constituting only a fifth of the sex chromosome. A possible explanation for this observation was provided by comparative transcriptomic analysis, which suggested that the number of sex-biased genes in *Ectocarpus* may be insufficient to drive SDR expansion, providing support for the sex antagonistic theory for SDR expansion. Both the male and female SDR haplotypes showed signs of degeneration despite the action of purifying selection during the haploid phase of the life cycle. Expression analysis data suggested that the genes that have escaped degeneration function preferentially during the haploid phase. Interest-

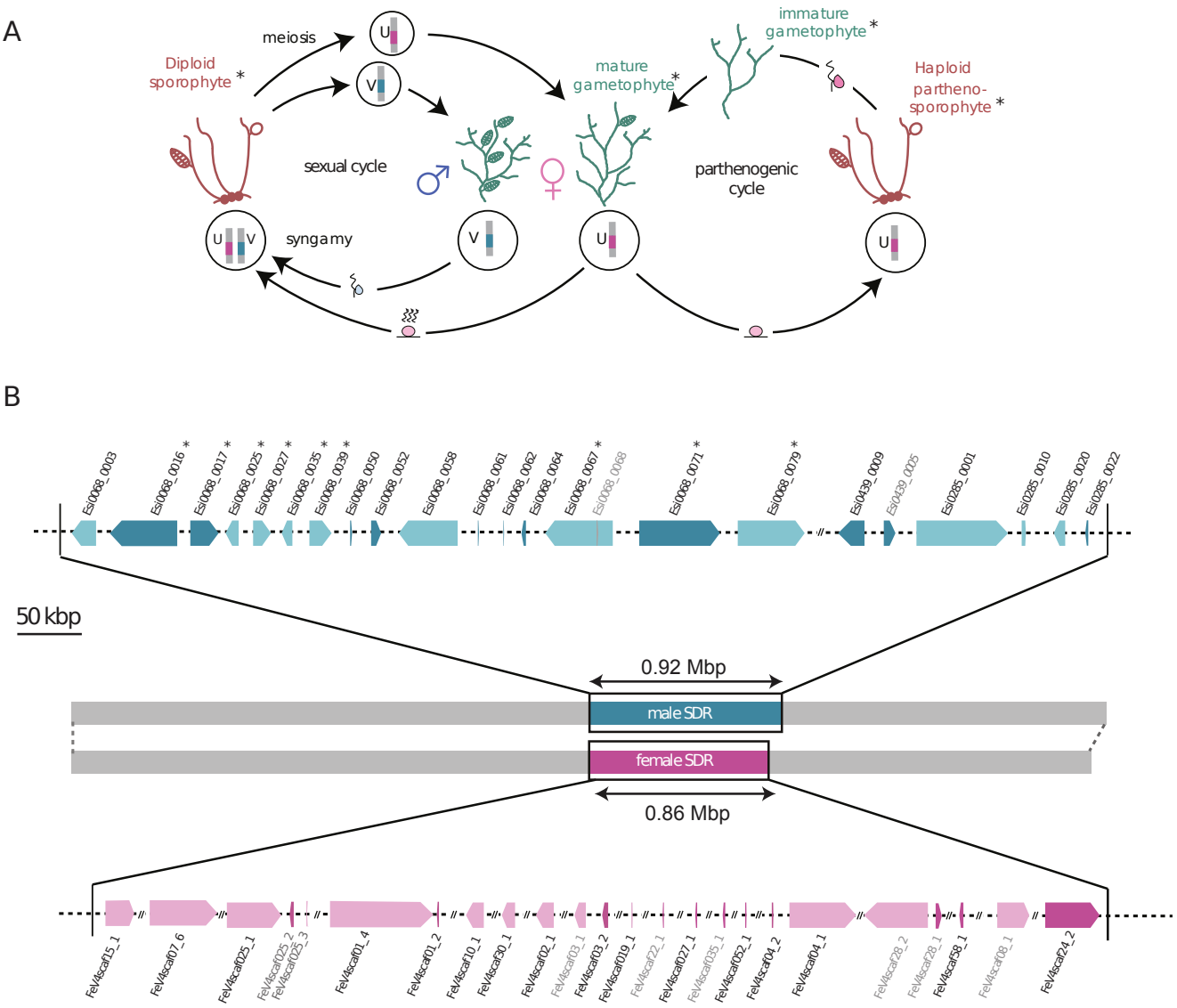


Fig. 1 The UV sex-determination system of the brown alga *Ectocarpus*. **(A)** Life cycle of *Ectocarpus* in culture. The sexual cycle (left side of panel) involves an alternation between the diploid sporophyte and haploid, dioecious (male and female) gametophytes. The sporophyte produces meio-spores through meiosis in the unilocular sporangia. The meio-spores are released and develop as gametophytes (each containing either a U or a V sex chromosome), which then produce gametes in plurilocular gametangia. Fusion of male and female gametes produces a zygote (containing both the U and the V sex chromosomes), which develops as a diploid sporophyte, completing the sexual cycle. Unfertilised gametes can enter an asexual parthenogenetic cycle by germinating without fusion to produce a partheno-sporophyte (right side of panel). The partheno-sporophyte produces spores in unilocular sporangia and these develop as gametophytes, completing the parthenogenetic cycle. Note that the haploid partheno-sporophytes and the diploid sporophytes do not express sex. The parthenogenetic cycle is only shown for a female but male gametes can also develop parthenogenetically. Life cycle stages used for the RT-QPCR analysis of SDR gene expression are marked with an asterisk. **(B)** Overview of the *Ectocarpus* male and female SDR haplotypes. Genes are indicated by arrows, the lighter colours corresponding to gametologues. Gene names (LocusIDs) are indicated, with pseudogenes in grey type and putative transposon remnants in grey italic. The relative sizes of the male and female SDR genes are indicated but they are not drawn to the same scale as the underlying scaffolds indicated by the dotted line and the scale bar. Asterisks indicate two clusters of genes that exhibited peak transcript abundance in sexually mature male gametophytes. Scaffolds are separated by double diagonal lines indicating that the relative positions of scaffolds within the SDR are unknown. Double-headed arrows indicate the sizes of the SDR haplotypes. The grey bars indicate the sex chromosomes. See also Figure S1.

Table 1 Statistics for several features of the male and female *Ectocarpus* SDR compared with the PAR and genome

	Male SDR	Female SDR	PAR	Genome
Total sequence (Mbp)	0.92	0.86	4.08	205.27
Genes (incl. pseudogenes)	23	24	239	16,015
Average gene length (bp)	20,021	13,987	7,189	6,828
Average CDS length (bp)	1,014	898	1,079	1,575
Average 3'UTR length (bp)	513	760	725	672
Average 5'UTR length (bp)	105	199	157	132
Average intron length (bp)	3,337	3,570	1,005	697
Average n. introns/gene	4.91	4.66	5.71	7.01
Gene density (genes/Mbp)	25.04	27.90	58.64	78.02
GC%	51.09	51.09	52.67	54.46
Alternative transcripts per gene	2.00	1.38	2.51	2.28

ingly, the male SDR haplotype was dominant over the female haplotype, even when the latter was present in two copies, suggesting that the V chromosome determines maleness, with femaleness being the default state when this chromosome is absent. A male-specific high mobility group (HMG) domain gene, which was most highly expressed during male fertility, was identified as a candidate male sex-determining gene. Analysis of the *Ectocarpus* SDR has underlined the universality of sex chromosome evolution across the eukaryotes and has provided important insights into sex chromosome evolution in UV sexual systems.

Results

Identification and characterisation of the *Ectocarpus* SDR

The *Ectocarpus* SDR was localised to linkage group 30 of the genetic map [24] by identifying scaffolds that showed male-specific hybridisation patterns in comparative genome hybridisation experiments [25]. These scaffolds were then located on the genetic map by genotyping the segregating population that was used to generate the map [24] with markers from the SDR scaffolds (Figure S1A, Table S1A-C). To confirm cosegregation of the SDR with sexual phenotype, 34 strains of known sex were genotyped with several sex locus markers, corresponding to both the male and female SDR haplotypes (Table S1D). In all cases the SDR genotype correlated with sexual phenotype confirming that this region is the sex-determining locus in *Ectocarpus*.

Further analysis of the segregation patterns of genetic markers corresponding to SDR scaffolds in a single family of 2000 siblings detected no recombination events (Figure S1B). The SDR therefore behaved as a discrete, non-recombining haplotype. This genetic analysis indicated that the male SDR extended over a region of approximately 920 kbp (Figure 1B, Table 1).

To characterise the female haplotype of the sex locus, we sequenced the genome of a female *Ectocarpus* strain that is closely related to the male sequenced strain (Figure S1A) [15]. Several strategies were used to identify candidate female SDR scaffolds (Supplemental Material, Tables S1E-G) and sex linkage was verified by genetic mapping (Table S1H). The cumulative size of the sex-linked scaffolds was 860 kbp, indicating that the male and female SDR haplotypes are of similar size (Figure 1B, Table 1). The SDR is flanked by two, large chromosomal domains corresponding to the pseudoautosomal region (PAR). Analysis of molecular marker segregation [24] indicated that the PAR, unlike the SDR, undergoes recombination during meiosis. For several parameters (gene density, intron length, percent GC content) the PAR had values that were intermediate between those of the autosomes and the SDR (Table 1).

Both the male and female SDR haplotypes are rich in transposable element sequences (Figure 2A) and gene poor compared to the autosomes (Table 1), features typical of non-recombining regions [1]. With only one exception (LTR transposons in the female SDR), all TE classes were more abundant in the SDR and the PAR than in autosomes, with the differences being particularly marked for both SDR haplotypes. Therefore, although the sex chromosome contains a higher percentage of transposon sequence throughout its length, there was clear evidence that diverse classes of transposon had accumulated to particularly high levels in the SDR. When individual classes of transposable elements were considered, retrotransposons (which represent the least abundant transposon class in the *Ectocarpus* genome as a whole) showed the most marked

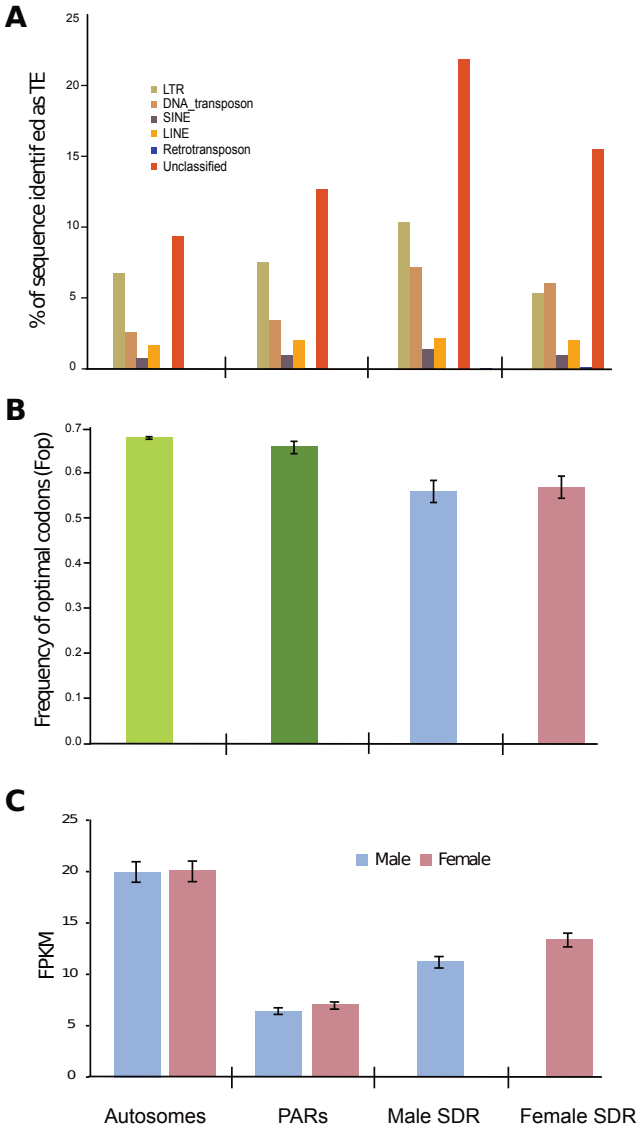


Fig. 2 Comparison of genomic features of the SDR, PAR and autosomes. (A) Percentage of DNA corresponding to different classes of transposable element (TE) in different genomic fractions. All comparisons were significantly different ($P<0.0001$). **(B)** Median frequency of optimal codons in coding regions of autosomal, PAR and male and female SDR genes. Error bars indicate 95% confidence intervals around the median. **(C)** Mean transcript abundance in sexually mature, male and female gametophytes for genes in different genome fractions, determined by RNAseq and expressed as fragments per kb of transcript per million fragments (FPKM) mapped. Error bars indicate 95% confidence intervals around the mean (1000 replications). See also Figure S2.

proportional enrichment in the SDR haplotypes compared to the autosomes (Figure S2A).

Sequences displaying intra chromosomal identities of 99.9% represent a large and distinct subset (30%) of the euchromatin of the human male-specific region of the Y-chromosome and this has been taken as evidence for a high level of gene conversion [5,26]. It was further suggested that gene conversion might "substitute" for inter-chromosomal recombination to some extent, counteracting the degenerative effects of reduced recombination within the SDR. In contrast, very little intra-haplotype sequence similarity was identified within either the male or the female *Ectocarpus* SDR haplotypes (Figures S2B and S2C). The total lengths of the repeated regions within the male and female SDR were only 1.14% and 2.16%, respectively. These low values suggest that intra-haplotype gene conversion has not been an important mechanism in the evolution of this SDR.

The male SDR haplotype contains 22 protein coding genes and one pseudogene, whereas 17 protein-coding genes and seven pseudogenes were found in the female haplotype (Figure 1B, Table S2). Eight of the female protein coding genes and three of the pseudogenes are homologous to male SDR sequences ("gametologues"), confirming a common autosomal origin for these two genomic regions. The classification of these genes as gametologues was supported by expression analysis, which showed that transcript abundances for gametologue pairs were strongly correlated (Figure S2D), and by their conserved intron/exon structures (Figure 3). The genes and pseudogenes that were only found in one (male or female) haplotype may have either been acquired since the divergence of the U and V regions or been lost by the counterpart haplotype. Seventeen of the male and female genes/pseudogenes that were found in only one haplotype had homologues outside the SDR (including, in two cases, genes on linkage group 30; Figure S3, Table S2). The presence of these autosomal homologues was consistent with a process of gene gain (i.e. via gene duplication events). The remaining five genes that were found in only one haplotype may represent cases of gene loss in the other haplotype, but they could also have resulted from gene relocation to the SDR. Testing these hypotheses will require comparison with a homologous gene from an outgroup species.

Genomic degeneration of the SDR region

Suppression of recombination across the SDR is expected to lead to genetic degeneration unless there is strong selection on gene function to counteract this effect. There are several indications that genetic degradation has occurred, at least to some degree, in the *Ectocarpus* SDR. We identified a set of optimal codons for *Ectocarpus* (Figure S2E and S2F). Selection on codon usage is known to be of weak intensity and particularly sensitive to loss of recombination [27,28]. The coding

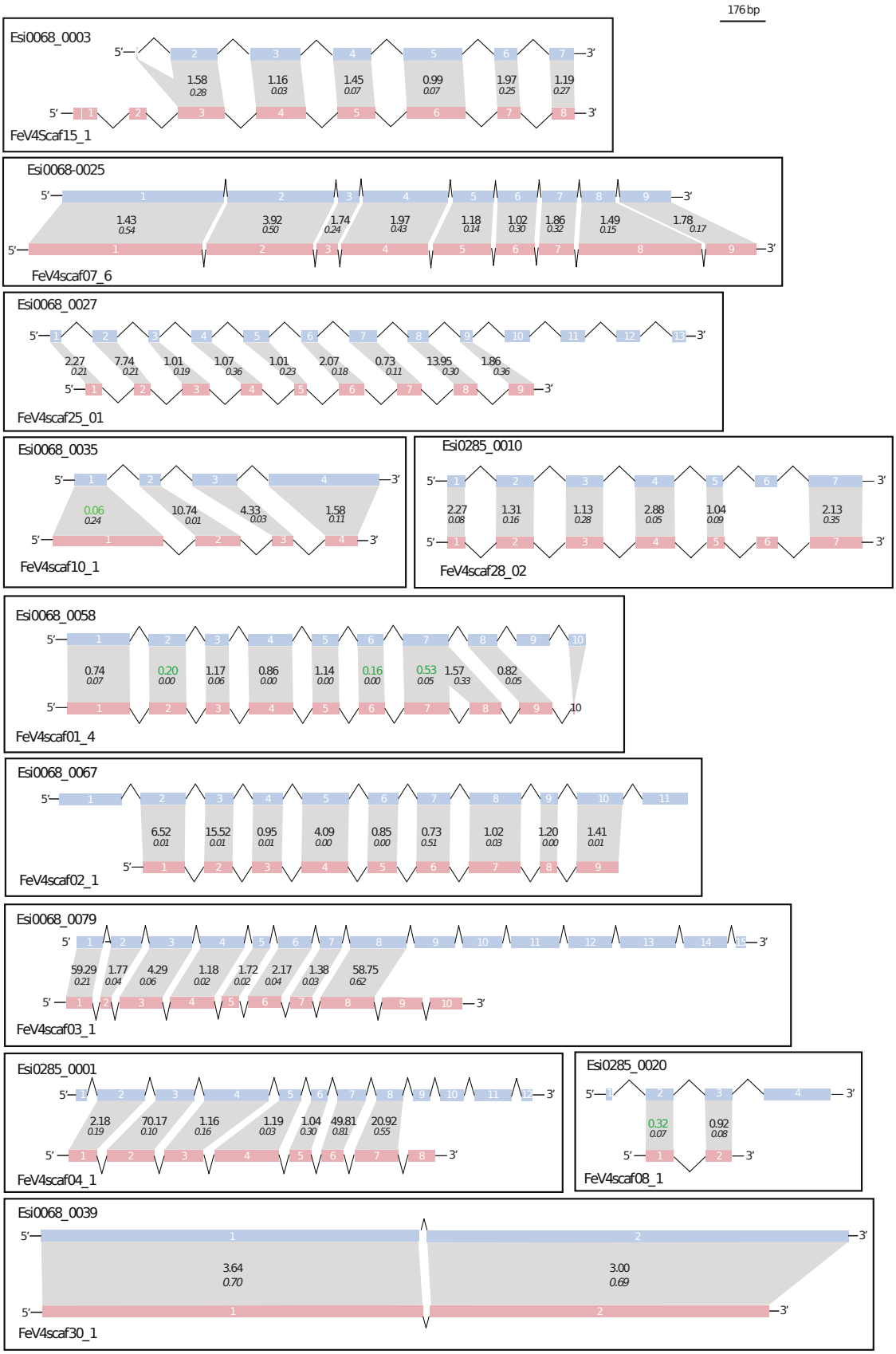


Fig. 3 Exon-by-exon analysis of synonymous site substitutions between gametologues. Male gametologues are shown in blue, female gametologues in red. Numbers in plain type indicate synonymous site substitution (dS) values between exons; numbers in italics indicate non-synonymous site substitution (dN) values. dS values in green indicate possible gene conversion events. See also Figure S3.

sequences of SDR genes exhibited a significant under-representation of optimal codons, suggesting maladapted codon usage (Figure 2B). Moreover, transcripts of SDR genes were less abundant on average than transcripts of non-SDR genes, which may reflect degradation of the promoter and cis-regulatory sequences of the SDR genes (Figure 2C). SDR genes were found to be much longer on average than genes elsewhere in the genome, due principally to the presence of longer introns (Table 1). This difference was partly explained by the presence of a larger amount of inserted transposable element DNA (Figures 2A and S2G), which is typical of non-recombining regions (although a direct role of TE insertions in degeneration still remains to be proven). Although these analyses detected genomic degeneration in the SDR, the overall degree of degeneration was modest compared to previously characterised systems [29], perhaps because both the U and V SDR haplotypes have essential functions during the haploid phase and are constantly exposed to selection (in contrast to Y or W chromosome genes which are always in an heterozygous context). An analysis of SDR gene expression supported this hypothesis: transcripts of SDR genes were consistently more abundant during the haploid compared with the diploid phase of the life cycle (Figure 4A, 4B, S4A-C). Another potential explanation for the limited degree of degeneration is that the SDR is small compared to most previously characterised systems and this may have limited the potential for Hill-Robertson interference among selected sites [30].

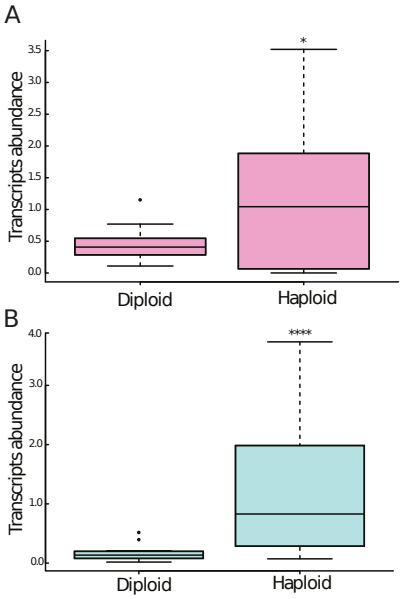


Fig. 4 SDR gene transcript abundance during the diploid and haploid phases of the *Ectocarpus* life cycle. Box plots showing the distributions of abundances of female **(A)** and male **(B)** SDR gene transcripts during the diploid (sporophyte) and haploid (gametophyte and partheno-sporophyte) phases of the *Ectocarpus* life cycle. $****$: $P < 0.001$ and $*$: $P \leq 0.05$. At least three independent biological replicates were used for each of the stages tested. See also Figure S4.

Predicted functions of SDR genes

Of the 11 genes that were found in the male but not the female SDR haplotype, one was of particular interest because it was predicted to encode a HMG domain protein (Figure S4D, Table S4A). This family of proteins has been implicated in sex or mating type determination in both vertebrates and fungi [31,32]. The SDR of the green alga *Volvox* also contains a HMG gene [12]. In addition, several of the genes that were found in both the male and female SDR haplotypes (gametologues) were predicted to encode potential signal transduction proteins (including a Ste20-like kinase, a casein kinase, a GTPase, a RING zinc finger protein and a MEMO domain protein; Table S2) and could potentially be involved in the regulation of sex determination.

An ancient sex-determining region

At the sequence level, the male and female haplotypes are extremely divergent (Figure 5A). No large blocks of sequence similarity were found and the only regions with a high level of similarity corresponded to gametologue exons (Figure 3). Most of the female haplotype scaffolds contained only one gametologue but in the single case where a scaffold contained two adjacent gametologues (FeV4scaf25_1 and FeV4scaf25_3) these genes were not contiguous on the male haplotype (68_27 and 68_68), suggesting that gene order is not conserved between the two haplotypes (Figure 1B and Table S2). This divergence suggests that the male and female haplotypes have been evolving independently over a long period. Two phylogenetic trees were constructed based on sequences of either an SDR or an autosomal gene from three *Ectocarpus* lineages and *Sphaerotrichia divaricata*, a distantly related brown algal species. The topology of the phylogenetic tree based on the autosomal gene was consistent with sequential speciation, with sequences from male and female strains of the same lineage grouping together (Figure 5B). In contrast, in the phylogenetic tree based on the SDR gene, sequences grouped together according to gender (Figure 5C). The presence of corresponding male and female sequences in *Ectocarpus* and *S. divaricata* suggests that the SDR stopped recombining at least 70 MYa [34]. The rate of synonymous substitution (dS) in the coding regions

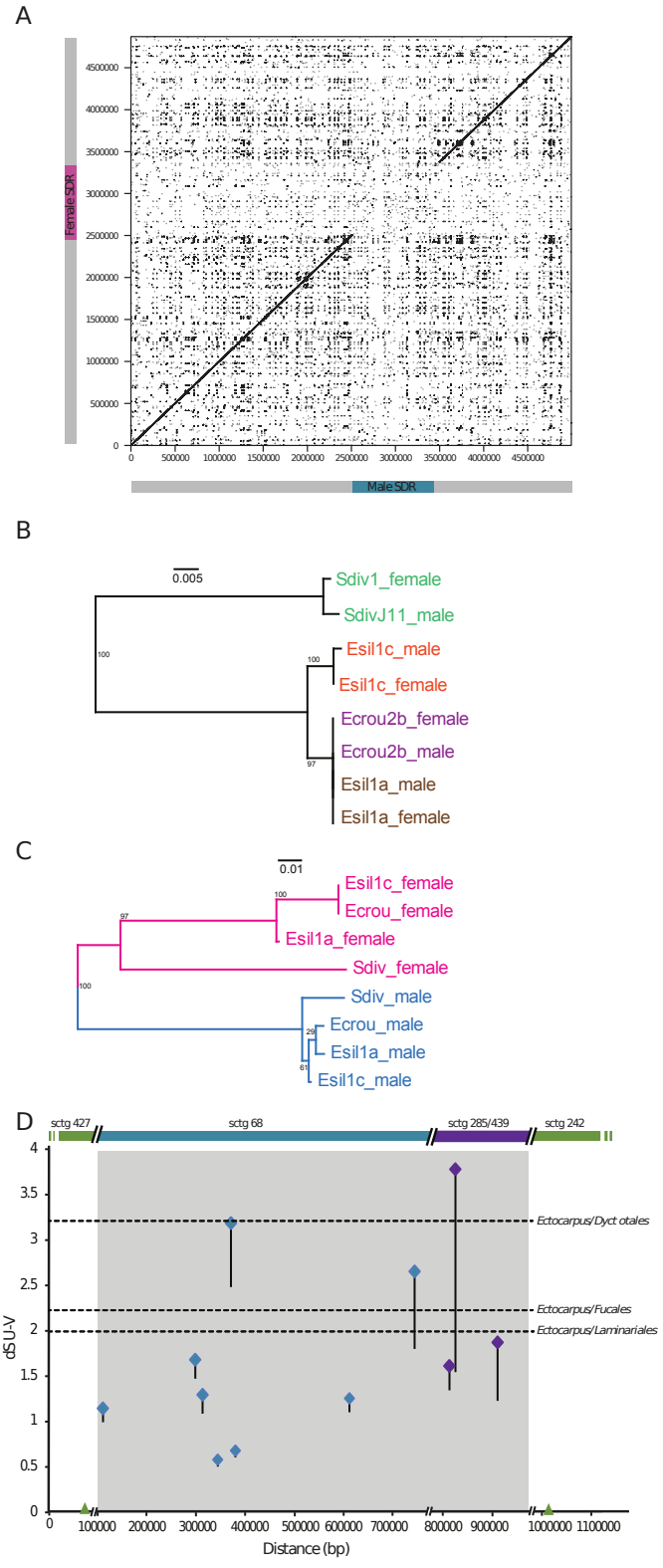


Fig. 5 The male and female SDR haplotypes are highly divergent. (A) Dotplot comparison of the *Ectocarpus* U and V chromosomes. The order and orientation of the SDR scaffolds is arbitrary. Note that the matches between corresponding exon regions of gametologues are not visible at this scale of analysis. **(B)** Unrooted neighbour-joining tree created in MEGA [33] using coding sequence data amplified from an autosomal gene. Distinct lineages are indicated by different colours. Samples correspond to three different *Ectocarpus* lineages and the related species, *Sphaerotrichia divaricate*. Lineage names and sex are indicated at the branch tips. Strains used are described in Table S1A. **(C)** Equivalent tree to that shown in **(B)** but for one gametologue pair. Pink and blue indicate sequences from male and female individuals respectively. **(D)** Plot of dS values of gametologue and PAR homologous pairs against gene distance, with gene order according to the male physical map. Blue and purple lozenges represent genes on the two male SDR scaffolds, sctg_68 and sctg_285and439, respectively. Green triangles at each end of the x-axis represent two flanking PAR genes. One-sided standard error bars represent half the standard error of the estimation. Double diagonal bars indicate that the orientation of the locus relative to the flanking PAR is not known. Dotted lines indicate mean levels of synonymous site divergence between *Ectocarpus* autosomal genes and autosomal genes of the species from the brown algal groups indicated. See also Figure S5.

of the 11 male and female gametologue pairs (Figure 5D) was used to independently evaluate the age of the SDR. The dS values for these gene pairs were compared with values for orthologous, autosomal gene pairs across eleven brown algal and diatom species for which divergence times had been estimated (Supplemental Material). The dS values for the SDR genes were remarkably high (mean value of 1.7, with most genes having dS ¿1) and comparisons with values obtained for the pairs of autosomal orthologues indicated that the male and female haplotypes of the SDR stopped recombining about 100-200 My ago (Figure S5). These analyses suggests that the *Ectocarpus* UV SDR is an old system comparable to the *Drosophila* (60 MY) [29] and mammalian (180 MY) [35,36] XY systems. Gametologue gene pair dS values were similar, with some exceptions, and it is possible that a single recombination suppression event gave rise to the *Ectocarpus* SDR.

When dS values were calculated on an exon-by-exon basis, individual exons with a markedly lower dS value than those of the other exons within the gametologue gene pair were identified for three of the 11 gametologue pairs (Figure 3). The presence of these rare variant exon pairs suggests that gene conversion events affecting individual exons or small gene regions have occurred since the divergence of the male and female SDR haplotypes.

Limited expansion of the *Ectocarpus* SDR

Given its age, and the prediction that an SDR should progressively enlarge over time to encompass a large part of its chromosome [1,37], it is remarkable that the *Ectocarpus* SDR has remained relatively small, accounting for only about one fifth of linkage group 30 and extending over less than a Mbp. It is possible that

the relatively small size of the SDR is related to the low level of sexual dimorphism in *Ectocarpus* as the recruitment of sexually antagonistic genes is believed to be an important driver of SDR expansion [1,37]. Moreover, sexually antagonistic polymorphisms are predicted to be less stable in haploid systems than in diploid systems because dominance effects in XX (or ZZ) individuals are expected to favour allele maintenance in the latter [38,39]. This effect may also limit expansion of the SDR by reducing the number of genes with sexually antagonistic polymorphisms available for recruitment into the SDR. Consistent with these hypotheses, comparison of the transcriptomes of male and female gametophytes indicated that only about 4% of *Ectocarpus* genes showed sex-biased expression (compared with 50-75% in *Drosophila* for example; [40,41]; Table S4B).

SDR gene expression and dominance

Quantitative PCR was used to measure the abundance of SDR gene transcripts in near-isogenic male and female strains (Figure S4A and S4B) at different stages of the life cycle (Figure 1A). While no clear pattern was observed for the female SDR genes (Figure S4B), transcripts of two thirds of the male SDR genes that were analysed were most abundant in mature gametophytes (Figure S4A) suggesting that these genes have a role in fertility. Interestingly, these putative fertility-related genes are organised into two clusters within the male SDR (Figure 1B). The first cluster includes the HMG domain gene, whose transcript was more than ten-fold more abundant in mature gametophytes than at the other stages assayed (Figure 6). The other fertility-induced genes included both additional male-specific genes (encoding conserved unknown proteins) and several gametologue pairs (predicted to encode a GTPase, a MEMO-like domain protein, a nucleotide transferase and a homoaconitate hydratase, for example) (Table S2).

Diploid gametophytes bearing both the male and the female SDR haplotypes (UV) can be generated artificially, and these individuals are always phenotypically male, indicating that the male haplotype is dominant [23,42]. This dominance relationship would be consistent with the existence of a male master sex-determining gene. To determine whether the dominance of the male haplotype is dose dependent, we used the life cycle mutant *ouroboros* [42] to construct seven independent triploid (UUV) gametophytes (Figure S1A, Table S1I). All seven triploids produced male gametes (as determined by genetic crosses with tester lines). Measurements of transcript abundances for 11 female

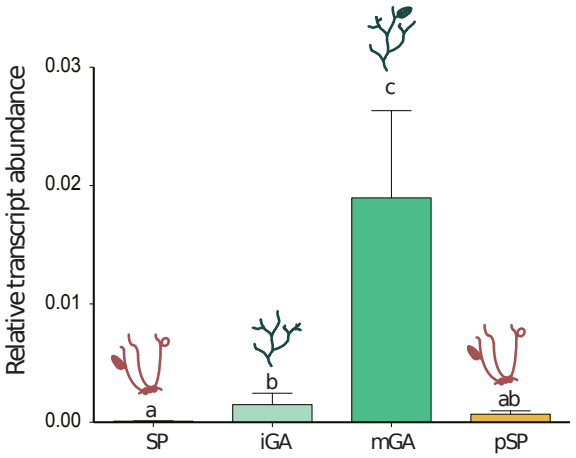


Fig. 6 The transcript of the HMG gene in the male SDR was most abundant in mature gametophytes. Transcript abundance determined by quantitative RT-PCR is given relative to the normalisation gene *EF1a*. Means of at least three biological replicate samples are shown. Error bars represent SE. SP, diploid sporophyte; iGA, immature gametophyte; mGA, mature gametophyte; pSP, partheno-sporophyte.

SDR genes did not detect a marked down-regulation of these genes in diploid heterozygous gametophytes compared to haploid gametophytes (Figure S4E and S4F). This suggests that the male haplotype does not silence female gene expression in this heterozygous context, although it was not possible to rule out that the expression of specific female haplotype genes was suppressed. It is possible, therefore, that gametophytes adopt the female developmental program by default, when the male SDR haplotype is absent.

Discussion

This study has demonstrated that sex is determined during the haploid phase of the brown alga *Ectocarpus* by a non-recombining region on linkage group 30 that extends over almost 1 Mbp. The male and female haplotypes of the SDR were of similar size but were highly diverged, the only detectable similarity being the presence of 11 gametologues, three of which were predicted to be pseudogenes in the female. Based on comparisons of these shared genes across diverse brown algal species, the SDR was estimated to be approximately 100-200 million years old. Compared with previously characterised systems [43], the *Ectocarpus* UV chromosomes can clearly be classed as an ancient (as opposed to a recently evolved) sex-determining system.

The brown algae belong to the Stramenopiles, which diverged from the lineages that led to green plants and animals more than a billion years ago [44]. This study therefore confirms that SDRs from diverse eukaryote

groups share a number of fundamental features such as stable maintenance of pairs of functional alleles (gametologues) over long periods of evolutionary time, suppressed recombination within the SDR, low gene density and accumulation of transposable elements. The presence of 11 gametologue pairs provided unambiguous evidence that the *Ectocarpus* UV pair is derived from an ancestral pair of autosomes, as has been observed for XY and ZW systems in animals and plants [1,37,7].

Analysis of the *Ectocarpus* SDR has also allowed a number of predictions that specifically concern UV sexual systems [8] to be tested. UV systems are not expected to exhibit the asymmetrical degeneracy of the sexual chromosomes (degeneracy of the Y and W chromosomes) observed in XY and ZW systems [29] and this supposition is supported by the similar sizes of the male and female SDR haplotypes in *Ectocarpus*. Gene degeneration was also observed to be modest within the *Ectocarpus* SDR and SDR gene transcripts tended to be more abundant during the haploid stages of the life cycle. These observations are consistent with purifying selection acting to maintain gene functionality during the haploid phase, when the U and V chromosomes are found in separate, male and female, organisms. Selection is indeed expected to be stronger during the haploid phase and to limit degeneration, as suggested for the relatively gene-dense V chromosome of *Marchantia* [11], another UV system, and by the low dN/dS ratios observed for sex-linked pollen-expressed genes in *Silene latifolia*, a plant with XY chromosomes [45]. The detection of modest levels of gene degeneration indicates that UV SDRs are nonetheless subject to the degenerating effects of suppressed recombination to some degree.

We cannot entirely rule out the possibility that the low gene content of the *Ectocarpus* U and V SDR haplotypes is a result of gene loss (as appears to be the case for the human Y chromosome, for example [29]), implying that the extant SDR is a remnant of a previously larger gene-rich region. However, gene loss is expected to be limited in UV systems because, assuming that this process will affect the two chromosomes symmetrically, each haplotype can only loose at most 50% of its essential genes. Gene losses of more than 50% would imply a significant proportion of non-essential genes in the nascent SDR and/or relocation of genes to other sites in the genome [8]. Future analysis of outgroup species in which the equivalent of the *Ectocarpus* sex chromosome has remained autosomal may help address this question.

Despite being ancient, the *Ectocarpus* SDR has remained relatively small, accounting for only about a fifth of the sex chromosome. Given the low level of sexual dimorphism in *Ectocarpus* and the small number of genes that show sex-biased expression, the small size of the SDR is consistent with the view that SDR expansion is driven by the evolution of genes with sexually antagonistic effects [1,46]. In a number of sex chromosome systems, the expansion of the non-recombining region of the Y (or W) has been shown to have proceeded through several events of recombination suppression, which have formed regions with different degrees of X-Y (or Z-W) divergence (evolutionary strata) [4,47] (reviewed in [1,43]). Although strata may be more difficult to detect in haploid systems (as both U and V can accumulate rearrangements), the lack of detectable strata is consistent both with this small SDR having arisen as the result of a single recombination suppressing event and with the conclusion that this region has experienced limited expansion. This does not mean that UV systems cannot acquire evolutionary strata, as recent evidence suggests the possible existence of at least two strata in the UV system of the bryophyte Ceratodon [13]. Note also that the *Ectocarpus* system provides independent evidence that the age of an SDR does not necessarily correlate perfectly either with its size or with the degree of heteromorphy (e.g. [48,49]).

In *Ectocarpus*, the male SDR haplotype was dominant over the female haplotype, even when two copies of the female haplotype were present. It is therefore possible that femaleness may simply be the default state, adopted when the male haplotype is absent. This situation is comparable to that observed in diverse animal, fungal and land plant sex-determination systems but differs from that observed with the UV systems of some mosses. In the latter, the male and female factors are co-dominant, leading to monoecy when both the male and female SDR haplotypes are present in the same gametophyte [33]. Functional differences can therefore be observed between different sex determination systems independent of the genetic nature of the system (XY, ZW or UV).

The male-specific HMG gene is a good candidate for the gene that determines maleness in *Ectocarpus*. If this can be confirmed experimentally it will raise important questions about the evolution of sex-determination gene networks across the eukaryote tree, suggesting shared or convergent mechanisms in brown algae, fungi and animals.

Experimental procedures

The raw sequence data generated in this study (Supplementary Information) have been submitted to Genbank with the accession number ERA209450.

Ectocarpus culture

Ectocarpus strains were cultured as described [50].

RNAseq transcriptome data

RNA-seq analysis was carried out to compare the abundances of gene transcripts in male and female mature gametophytes. Synchronous cultures of gametophytes of the near-isogenic male and female lines Ec603 and Ec602 (see Table S1A, Figure S1) were prepared under standard conditions [50] and frozen at maturity. Total RNA was extracted from 2 bulks of 400 male individuals and 2 bulks of 400 female individuals (2 biological replicates for each sex) using the Qiagen Mini kit (`http://www.qiagen.com`) as previously described [42]. For each replicate, RNAs were quantified, cDNAs for transcriptome analysis were dT primed, fragmented, cloned, and sequenced by Fasteris (CH-1228 Plan-les-Ouates, Switzerland). We used both de novo assembly (Trinity) (r2012-01-25) [51] and TopHat (v2.0.3) [52, 53] and Cufflinks (v2.0.2) [53,54] algorithms. Statistical testing for sex biased gene expression was performed using DEseq [55].

Identification and mapping of the male SDR

A comparative genome hybridisation approach [25] identified several regions of the genome exhibiting polymorphisms between male (Ec32) and female (Ec568) strains. Primers were developed for these putative sex-linked regions and mapping was performed by genotyping the 60 individuals of the mapping population [24]. Details of the PCR conditions are given in the Supplemental Information. The approaches used to improve the assembly of the male SDR and the verification of the completeness of the male SDR using an RNA-seq based method are explained in detail in the Supplemental Information section.

Recombination analysis

Recombination between sex locus markers was analysed using a large segregating family of 2000 meiotic individuals (Figure S1) derived from a cross between the male line Ec494 [42] and the female outcrossing line Ec568 [24].

Sequencing of a female strain and identification and assembly of the female SDR

The genome of the female strain Ec597 (Table S1A, Figure S1A) was sequenced using a whole genome shotgun strategy that involved the implementation of both Illumina HiSeq 2000 technology and Roche 454 pyrosequencing. `Velvet` (version 1.1.05) was used to run several assemblies during the sequencing process, including the V3 assembly (which used all the pair-end (PE) reads and reads from one of the mate-pair libraries) and the final V4 assembly with the complete read dataset (Table S1E). An independent de novo assembly was also carried out with the `CLC` assembler (`http://www.clcbio.com/products/clc-assembly-cell`) using only the PE Illumina data.

Female SDR scaffolds were identified using two different approaches. First the deduced protein sequences of male SDR genes (all annotated genes on the two male SDR scaffolds sctg_68 and sctg_285and439) were blasted against the female genome assembly. Fourteen candidate female SDR scaffolds were identified in the V4 assembly using this approach. The second approach employed RNA-seq transcriptome data. All putative female specific scaffolds were verified by PCR using between eight and 57 individuals. Several approaches were used to improve the assembly of the female SDR. Details are given in the Supplemental Information section.

Annotation of SDR scaffolds

The male SDR scaffolds had been annotated as part of the *Ectocarpus* genome project [15] but the gene models were considerably improved by integrating transcript information derived from the RNA-seq analysis carried out as part of this study and using comparisons of male and female gametologue gene models. The updated gene models can be accessed at the OrcAE database (`http://bioinformatics.psb.ugent.be/orcae/overview/Ectsi`) [56]. The female SDR scaffolds were annotated de novo by running the gene prediction program `EuGène` [57], which incorporated the signal prediction program `SpliceMachine` [58], using the optimised Markov models and `SpliceMachine` splice site predictions derived previously for the male genome sequence [15]. Gene prediction incorporated extrinsic information from mapping of the RNA-seq data onto the female-specific scaffolds. Both male and female SDR gene models were manually curated using the raw, mapped RNA-seq data, `Cufflinks` and `Trinity` transcript predictions and comparisons between the male and female haplotypes.

Pseudogenes were identified manually by comparing SDR sequences with genes in the public databases. An additional screen for pseudogenes was carried out by blasting male protein sequences against the genomic sequence of the female SDR and vice versa. All sequences that had been annotated as "gene" or "TE" were excluded from this latter analysis using `Makseq` and `RepeatMasker` respectively.

Homologous genes present in both the male and female haplotypes of the SDR were considered to be gametologues if they were detected as matches in a reciprocal `Blastp` search against the SDR scaffolds (E value cutoff: $1e^{-4}$). The same criterion was used to identify homologues of SDR genes located outside the SDR (Table S2).

Identification of transposons and other repeated sequences in the SDR

An *Ectocarpus*-specific TE-library (described in [15]), which had been compiled with `Repet` [59], was used to annotate SDR transposons. TEs were also annotated by running the de novo annotation software `Repclass` [60] with default parameters. See the Supplemental information section for details.

Intra-haplotype sequence similarity

Analyses of sequence similarity within the male and female SDR haplotypes were performed using a custom Perl code [5]. By default, the threshold for sequence identity was fixed to 97%. When the threshold was reduced to 50%, the same result was obtained.

Quantitative reverse transcriptase PCR analysis of SDR gene transcript abundances during the *Ectocarpus* life cycle

The abundance of male and female SDR gene transcripts during the *Ectocarpus* life cycle was assessed by RT-QPCR. Primer pairs were designed to amplify regions of the 3'UTR or the most 3' exon of the gene to be analysed (Table S4C). In silico virtual PCR amplifications were carried out using the e-PCR program [61] and both the male and female genome sequences to check the specificity of oligonucleotide pairs. RT-QPCR analysis was carried out for 13 male SDR genes and 11 female SDR genes (Figure S4A and S4B). The remaining SDR genes could not be analysed either because they had very small exons, which posed a problem for primer design, or it was not possible to obtain a single amplification product. RNA extraction and RT-QPCR were performed as previously [42].

Construction of phylogenetic trees for an SDR and an autosomal gene

Exon sequences from an SDR and an autosomal gene were amplified from three *Ectocarpus* lineages (order Ectocarpales) and a related brown alga *Sphaerotrichia divaricata* (C. Agardh) Kylin. For the SDR gene, an exon region was amplified for the gametologue pair Esi68_0003 (male) and FeV4scaf15_1 (female). Sequence data from the ITS2 nuclear autosomal region was obtained for the same samples. Sequences were edited using the `Codon Code` sequence aligner and aligned with `Muscle` in the program `Seaview` [62]. Evolutionary history was inferred using both the Neighbour-joining (Figure 5C) and PhyML method implemented in `MEGA5` [63] with the same topology resolved by both methods. The strains and lineages used are described in Table S1A and the primers are described in Table S3.

Synonymous divergence

Alignments of protein sequences between gametologous gene pairs were performed in `Seaview` using `Muscle` with default parameters. Regions with poor alignments were further analysed with `Gblocks` [64]. The aligned protein sequences were then back-translated to coding sequence and synonymous divergence (dS) was calculated using `Codeml` within the suite of programs in `PAML` version 4 [65].

Estimating the age of the *Ectocarpus* SDR

Coding sequence data from 65 stramenopile species including two diatoms were obtained from the Hogenom database version 6 and from Genbank [66]. Homologous genes were identified using a clustering approach. Orthologous sequences were identified and checked using phylogenetic information (described in Supporting Information). Coding sequences from other Phaeophyceae species were added to the cluster data and further data cleaning was carried out so that only orthologous sequences were retained, as described in Supporting Information. A pairwise alignment of the *Ectocarpus* genes with all of the identified orthologous genes from each cluster was then carried out using Prank [67], and alignments were improved using `Gblocks` [64,65]. The programs Codeml and Yn00 from `PAML` version 4 [65] were then run on each gene pair in order to calculate pairwise

dS values. The resulting dS values were plotted against the divergence times estimated by Silberfeld et al. [34] and Brown and Sorhannus [68].

Codon usage analysis

A set of 27 optimal codons was identified by comparing the codon usage of highly expressed genes (ribosomal genes) with the rest of the genome using the multivariate approach described in Charif *et al.* [69]. Fop values were correlated with RNA-seq expression levels (Figure S2E and S2F).

Sex-determination in strains carrying different numbers of U and V chromosomes

Polyploid gametophytes were constructed using the *ouroboros* mutant [42] (Figure S1A). Details of genetic crosses and ploidy verification are given in Supporting Information.

Acknowledgements The authors wish to thank Bernard Billoud and Veronique Storm for advice on the statistical analysis, Aurélie Kapusta for help with Repclass, Emmanuelle Lerat for explanations about TE librairies, Thomas Bigot and Florent Lassalle for help with `TPMS`, Catherine Leblanc, Florian Weinberger, Gareth Pearson, Olivier de Clerk for sharing unpublished 454 and RNA-seq data and Helen Skaletsky for help with intra-chromosomal similarity analyses.

This work was supported by the Centre National de la Recherche Scientifique, the Agence Nationale de la Recherche (Project Sexseaweed), the University Pierre and Marie Curie Emergence program, the Interreg program France (Channel)-England (project Marinexus) and the Interreg IVB EnAlgae project. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Charlesworth D, Charlesworth B, Marais G (2005) Steps in the evolution of heteromorphic sex chromosomes. Heredity (Edinb) 95: 118–128.
2. Jordan CY, Charlesworth D (2012) The potential for sexually antagonistic polymorphism in different genome regions. Evolution 66: 505–516.
3. Ironside JE (2010) No amicable divorce? challenging the notion that sexual antagonism drives sex chromosome evolution. Bioessays 32: 718–726.
4. Lahn BT, Page DC (1999) Four evolutionary strata on the human x chromosome. Science 286: 964–967.
5. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.
6. Lemaitre C, Braga MD, Gautier C, Sagot MF, Tannier E, et al. (2009) Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. Genome biology and evolution 1: 56.
7. Wang J, Na JK, Yu Q, Gschwend AR, Han J, et al. (2012) Sequencing papaya x and yh chromosomes reveals molecular basis of incipient sex chromosome evolution. Proc Natl Acad Sci U S A 109: 13710–13715.
8. Bull JJ, et al. (1983) Evolution of sex determining mechanisms. The Benjamin/Cummings Publishing Company, Inc.
9. Bachtrog D, Kirkpatrick M, Mank JE, McDaniel SF, Pires JC, et al. (2011) Are all sex chromosomes created equal? Trends Genet 27: 350–357.
10. Bachtrog D (2011) Plant sex chromosomes: a non-degenerated y? Curr Biol 21: R685–R688.
11. Yamato KT, Ishizaki K, Fujisawa M, Okada S, Nakayama S, et al. (2007) Gene organization of the liverwort y chromosome reveals distinct sex chromosome evolution in a haploid system. Proc Natl Acad Sci U S A 104: 6472–6477.
12. Ferris P, Olson BJSC, De Hoff PL, Douglass S, Casero D, et al. (2010) Evolution of an expanded sex-determining locus in volvox. Science 328: 351–354.
13. McDaniel SF, Neubig KM, Payton AC, Quatrano RS, Cove DJ (2013) Recent gene-capture on the uv sex chromosomes of the moss ceratodon purpureus. Evolution 67: 2811–2822.
14. Peters AF, Marie D, Scornet D, Kloareg B, Mark Cock J (2004) Proposal of ectocarpus siliculosus (ectocarpales, phaeophyceae) as a model organism for brown algal genetics and genomics. Journal of Phycology 40: 1079–1088.
15. Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, et al. (2010) The ectocarpus genome and the independent evolution of multicellularity in brown algae. Nature 465: 617–621.
16. Menkis A, Jacobson DJ, Gustafsson T, Johannesson H (2008) The mating-type chromosome in the filamentous ascomycete neurospora tetrasperma represents a model for early evolution of sex chromosomes. PLoS Genet 4: e1000030.
17. Billiard S, López-Villavicencio M, Devier B, Hood ME, Fairhead C, et al. (2011) Having sex, yes, but with whom? inferences from fungi on the evolution of anisogamy and mating types. Biol Rev Camb Philos Soc 86: 421–442.
18. Hood ME, Petit E, Giraud T (2013) Extensive divergence between mating-type chromosomes of the anther-smut fungus. Genetics 193: 309–315.
19. Berthold G (1881) Die geschlechtliche fortpflanzung der eigentlichen phaeosporeen. Mitt Zool Stat Neapel 2: 401–413.
20. Van den Hoek C, Mann D, Jahns H (1995) Algae: an introduction to phycology (pp. 165–218). Cambridge University Press.
21. Evans L (1963) A large chromosome in the laminarian nucleus. Nature 198: 215.
22. Lewis RJ (1996) Chromosomes of the brown algae. Phycologia 35: 19–40.
23. Muller D (1975) Sex expression in aneuploid gametophytes of the brown alga ectocarpus siliculosus (dillw.) lyngb. Archiv fur Protistenkunde 117: 297–302.
24. Heesch S, Cho GY, Peters AF, Le Corguillé G, Falentin C, et al. (2010) A sequence-tagged genetic map for the brown alga ectocarpus siliculosus provides large-scale assembly of the genome sequence. New Phytol 188: 42–51.
25. Dittami SM, Proux C, Rousvoal S, Peters AF, Cock JM, et al. (2011) Microarray estimation of genomic interstrain variability in the genus ectocarpus (phaeophyceae). BMC Mol Biol 12: 2.

26. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, et al. (2003) Abundant gene conversion between arms of palindromes in human and ape y chromosomes. Nature 423: 873–876.
27. Bachtrog D (2003) Adaptation shapes patterns of genome evolution on sexual and asexual chromosomes in drosophila. Nat Genet 34: 215–219.
28. Bartolomé C, Charlesworth B (2006) Evolution of amino-acid sequences and codon usage on the drosophila miranda neo-sex chromosomes. Genetics 174: 2033–2044.
29. Bachtrog D (2013) Y-chromosome evolution: emerging insights into processes of y-chromosome degeneration. Nat Rev Genet 14: 113–124.
30. Hill WG, Robertson A (1966) The effect of linkage on limits to artificial selection. Genet Res 8: 269–294.
31. Foster JW, Brennan FE, Hampikian GK, Goodfellow PN, Sinclair AH, et al. (1992) Evolution of sex determination and the y chromosome: Sry-related sequences in marsupials. Nature 359: 531–533.
32. Idnurm A, Walton FJ, Floyd A, Heitman J (2008) Identification of the sex genes in an early diverged fungus. Nature 451: 193–196.
33. Allen CE (1935) The genetics of bryophytes. The Botanical Review 1: 269–291.
34. Silberfeld T, Leigh JW, Verbruggen H, Cruaud C, de Reviers B, et al. (2010) A multi-locus time-calibrated phylogeny of the brown algae (heterokonta, ochrophyta, phaeophyceae): Investigating the evolutionary nature of the "brown algal crown radiation". Mol Phylogenet Evol 56: 659–674.
35. Potrzebowski L, Vinckenbosch N, Marques AC, Chalmel F, Jégou B, et al. (2008) Chromosomal gene movements reflect the recent origin and biology of therian sex chromosomes. PLoS Biol 6: e80.
36. Veyrunes F, Waters PD, Miethke P, Rens W, McMillan D, et al. (2008) Bird-like sex chromosomes of platypus imply recent origin of mammal sex chromosomes. Genome Res 18: 965–973.
37. Bergero R, Charlesworth D (2011) Preservation of the y transcriptome in a 10-million-year-old plant sex chromosome system. Curr Biol 21: 1470–1474.
38. Rice WR (1984) Sex chromosomes and the evolution of sexual dimorphism. Evolution 38: 735–742.
39. Fry JD (2010) The genomic location of sexually antagonistic variation: some cautionary comments. Evolution 64: 1510–1516.
40. Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet 8: 689–698.
41. Assis R, Zhou Q, Bachtrog D (2012) Sex-biased transcriptome evolution in drosophila. Genome Biol Evol 4: 1189–1200.
42. Coelho SM, Godfroy O, Arun A, Le Corguillé G, Peters AF, et al. (2011) Ouroboros is a master regulator of the gametophyte to sporophyte life cycle transition in the brown alga ectocarpus. Proc Natl Acad Sci U S A 108: 11518–11523.
43. Bergero R, Charlesworth D (2009) The evolution of restricted recombination in sex chromosomes. Trends Ecol Evol 24: 94–102.
44. Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D (2004) A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol 21: 809–818.
45. Chibalina MV, Filatov DA (2011) Plant y chromosome degeneration is retarded by haploid purifying selection. Curr Biol 21: 1475–1479.
46. Qiu S, Bergero R, Charlesworth D (2013) Testing for the footprint of sexually antagonistic polymorphisms in the pseudoautosomal region of a plant sex chromosome pair. Genetics 194: 663–672.
47. Ellegren H, Carmichael A (2001) Multiple and independent cessation of recombination between avian sex chromosomes. Genetics 158: 325–331.
48. Stöck M, Horn A, Grossen C, Lindtke D, Sermier R, et al. (2011) Ever-young sex chromosomes in european tree frogs. PLoS Biol 9: e1001062.
49. Vicoso B, Kaiser VB, Bachtrog D (2013) Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. Proc Natl Acad Sci U S A 110: 6453–6458.
50. Coelho SM, Scornet D, Rousvoal S, Peters NT, Dartevelle L, et al. (2012) How to cultivate ectocarpus. Cold Spring Harb Protoc 2012: 258–261.
51. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from rna-seq data without a reference genome. Nat Biotechnol 29: 644–652.
52. Trapnell C, Pachter L, Salzberg SL (2009) Tophat: discovering splice junctions with rna-seq. Bioinformatics 25: 1105–1111.
53. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nat Protoc 7: 562–578.
54. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515.
55. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.
56. Sterck L, Billiau K, Abeel T, Rouzé P, Van de Peer Y (2012) Orcae: online resource for community annotation of eukaryotes. Nat Methods 9: 1041.
57. Foissac S, Gouzy J, Rombauts S, Mathé C, Amselem J, et al. (2008) Genome annotation in plants and fungi: Eugene as a model platform. Current Bioinformatics 3: 87–97.
58. Degroeve S, Saeys Y, De Baets B, Rouzé P, Van de Peer Y (2005) Splicemachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 21: 1332–1338.
59. Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6: e16526.
60. Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009) Exploring repetitive dna landscapes using repclass, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1: 205–220.
61. Schuler GD (1997) Sequence mapping by electronic pcr. Genome Res 7: 541–550.
62. Gouy M, Guindon S, Gascuel O (2010) Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27: 221–224.
63. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.
64. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.

65. Yang Z (2007) Paml 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.
66. Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, et al. (2009) Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10 Suppl 6: S3.
67. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 102: 10557–10562.
68. Brown JW, Sorhannus U (2010) A molecular genetic timescale for the diversification of autotrophic stramenopiles (ochrophyta): substantive underestimation of putative fossil ages. PLoS One 5.
69. Charif D, Thioulouse J, Lobry JR, Perrière G (2005) Online synonymous codon usage analyses with the ade4 and seqinr packages. Bioinformatics 21: 545–547.

Supporting Information

Supplementary methods

Ectocarpus strains and culture

Table S1A lists the strains used in this study. See also Figure S1A showing the pedigree of the strains used. The female strain used for genome sequencing, Ec597, was derived by crossing the ouroboros mutant Ec494 [1] with a female strain Ec419 (itself derived from a cross between a female strain Ec25 and the immediate upright mutant Ec137). Ec25, Ec137 and the male genome sequenced strain Ec32 [2] are all meiotic offspring of a field sporophyte, Ec17, collected in 1988 in San Juan de Marcona, Peru [3]. Ec494 is a UV-mutagenised descendant of Ec32. Two near-isogenic male and female inbred lines Ec601 and Ec602 were derived by repeated crossing of male and female progeny of Ec25 and Ec137 for eight generations. Ec569 was derived by crossing the male genome sequenced strain Ec32 with a female outcrossing line Ec568 from Arica in northern Chile [4]. Ec702 was derived by crossing the ouroboros mutant Ec494 with Ec568. *Ectocarpus* strains were cultured as described [5].

Generation of RNAseq transcriptome data for male and female gametophytes

RNAseq analysis was carried out as described in the main text. All material was checked under both a binocular microscope and at higher magnification to confirm fertility (presence of plurilocular gametangia) prior to total RNA extraction. RNAseq reads (100 bp read length, between 21 and 32 million reads per replicate) were trimmed and filtered with the FASTX toolkit (v0.0.13) using a quality threshold of 25 (base calling) and a minimal size limit of 60 nucleotides. We only retained reads in which more than 75% of nucleotides had a minimal quality threshold of 20.

Two assembly methods were used: de novo assembly using Trinity (r2012-01-25) [6] and reference-based assembly using the male genome sequence (Ec32) and a combination of the TopHat (v2.0.3) [7, 8] and Cufflinks (v2.0.2) [8, 9] algorithms. Each of the four datasets (duplicate male and female samples) was treated separately. Based on the Cufflinks analysis and statistical testing using DEseq [10], only about 4% of the genes expressed during this stage of the life cycle exhibited statistically significant sex-biased expression (Table S4B). The RNAseq data was also exploited to identify sex- linked genomic scaffolds (see below). Further details about the RNAseq analysis will be published elsewhere.

Identification and mapping of the male SDR

An *Ectocarpus* gene expression microarray based on EST sequences from the genome-sequenced strain (Ec32) has been used to carry out comparative genome hybridizations for several *Ectocarpus* strains [11]. This procedure identified regions of the genome that exhibited significant differences between the Ec32 (male) and Ec568 (female) strains, including two scaffolds that were highly polymorphic along their entire lengths (sctg_68 and sctg_439) and several scaffolds that exhibited polymorphism along only part of their length (sctg_15, sctg_285, sctg_474, sctg_595 and sctg_598).

PCR markers (Table S1C) were designed for each polymorphic scaffold using Primer3 [12] and in silico virtual PCR amplifications were carried out using the e-PCR program [13] to identify oligonucleotide pairs that were predicted to amplify a single region of the male genome. Sex linkage was initially tested using genomic DNA from four male and four female *Ectocarpus* strains (whose phenotypic sex had been

determined by crosses with reference strains). Markers that exhibited sex linkage in this preliminary test were then located on the *Ectocarpus* genetic map by genotyping the 60 individuals of the mapping population [4]. Markers were added to the genetic map using MAPMAKER [14]. Marker amplification PCRs were performed using the Promega PCR kit GoTaq® Flexi DNA Polymerase in a total volume of 20 μL containing 2 μL of 1:10 diluted DNA, primers at 100 nM, buffer at 1X, MgCl2 at 2mM, dNTPs at 200 μM and 0.5 units of Taq DNA Polymerase. The thermal profile included an initial denaturation step at 95°C for two minutes followed by 30 cycles of 95°C for 30 seconds, 60°C for 30 seconds and 72°C for 30 seconds. A final polymerisation step was carried out at 72°C for two minutes. For each PCR, we multiplexed an internal positive control (R26S) [15] to verify the efficiency of the PCR amplification. Using this approach, sctg_68, sctg_285 and sctg_439 were mapped to the *Ectocarpus* SDR. Scaffolds sctg_15, sctg_474, sctg_595 and sctg_598 were not sex linked.

RNAseq-based search for additional male SDR scaffolds

The RNAseq transcriptome data for male and female gametophytes was used to assess completeness of the male SDR. First, all scaffolds that encoded transcripts with sex-biased expression (FPKM>1 in male samples and FPKM<1 in female samples) were identified. These could have corresponded either to male SDR scaffolds or to scaffolds from other regions of the genome that carried genes with male-biased expression patterns. To eliminate the latter, we examined the context of the genes exhibiting male-biased expression and eliminated cases where these genes were surrounded by non-sex-biased genes. This approach yielded nine new candidate sex locus scaffolds (i.e. in addition to sctg_68, sctg_285 and sctg_439) but comparisons with the female genome revealed that eight of these scaffolds exhibited no significant polymorphism between sexes and genetic mapping of the remaining scaffold showed that it was not sex linked. The results of this analysis therefore suggested that the reconstruction of the male haplotype of the SDR was essentially complete.

Approaches used to improve the assembly of the male SDR haplotype

Although efforts to produce a large insert BAC library for *Ectocarpus* have been unsuccessful, a library of mini-BACs with an insert size of about 15 kbp has been generated [2]. Using BAC end sequence data, this library was screened in silico for BACs that spanned two of the three male SDR scaffolds. One BAC was detected (KY0FIPA75YN02) that matched ends of both sctg_285 and sctg_439, allowing these two scaffolds to be fused into a single scaffold (Table S1C). No BACs were detected that linked sctg_68 to either sctg_285 or sctg_439. The colinearity of sctg_285 and sctg_439 was confirmed independently by amplification of cDNA corresponding to the gene Esi0285_0001, which spans the two scaffolds, using oligonucleotide primers that corresponded to exons at the two adjacent ends of the scaffolds. The scaffold formed by assembling sctg_285 and sctg_439 is referred to as sctg_285and439.

Recombination analysis

Recombination between four sex locus markers corresponding to scaffolds sctg_68 and sctg_285and439 (Table S1B) was analysed using a large segregating family of 2000 meiotic individuals (Figure S1A) derived from a cross between the male line Ec494 [1] and the female outcrossing line Ec568 [4]. Between 200 and 1500 ng/μL of DNA was extracted from 100 mg of tissue (fresh weight) from each individual in the population using the Nucleospin® Multi-96 plant kit (Macherey-Nagel) according to the manufacturer's protocol and diluted 1:10. PCR reactions were performed using the Promega PCR kit GoTaq® Flexi DNA Polymerase in a total volume of 20 μL containing 2 μL of 1:10 diluted DNA,

oligonucleotides at 100 nM, buffer at 1X, MgCl2 at 2mM, dNTPs at 200 μM and 0.5 units of Taq DNA Polymerase. The thermal profile included an initial denaturation step at 95°C for two minutes followed by 30 cycles of 95°C for 30 seconds, 60°C for 30 seconds and 72°C for 30 seconds. A final polymerisation step was carried out at 72°C for two minutes. An internal positive control (R26S) [15] was used to verify the efficiency of PCR amplification. No recombination was detected between any of the markers located within the male SDR (Figure S2).

Sequencing of a female genome

The genome of the female strain Ec597 (Table S1A, Figure S1) was sequenced using a whole genome shotgun strategy using both Illumina HiSeq 2000 technology and Roche 454 pyrosequencing. One paired-end (PE) library with fragment sizes of about 180 bp, and two mate-pair (MP) libraries with insert sizes of 10 kbp were constructed for the Illumina sequencing. In total 18.1 Gbp data of paired-end reads of 104 bp and 9.9 Gbp of mate-pair reads of 51 bp were generated from these libraries.

Velvet (version 1.1.05) was used to run several assemblies during the sequencing process, including the V3 assembly (which used all the PE reads and reads from one of the MP libraries) and the final V4 assembly with the complete read dataset (Table S1E). The V3 assembly is the raw Velvet output, launched with a kmer value of 45. The V4 assembly was generated using Velvet with a kmer value of 51 follow by a step of gap closing using the tools provided with the SOAP de novo assembler (GapCloser). An independent de novo assembly was also carried out with the CLC assembler (`http://www.clcbio.com/products/clc-assembly-cell`) using only the paired end Illumina data.

Identification of scaffolds corresponding to the female haplotype of the SDR

Female SDR scaffolds were identified using two different approaches. First the deduced protein sequences of male SDR genes (all annotated genes on the two male SDR scaffolds sctg_68 and sctg_285and439) were blasted against both the V3 and V4 versions of the female genome assembly to detect scaffolds carrying female alleles (gametologues) of the male SDR genes. Fourteen candidate female SDR scaffolds were identified in the V4 assembly using this approach. The second approach employed the RNAseq transcriptome data. The two sets of female transcripts constructed by Trinity using the replicate female RNAseq datasets were independently compared with the male and female genome assemblies by local alignment using Blast. Transcripts that aligned with the female genome but not with the male genome were retained (E value cut-off: $1.e^{-4}$). The two sets of female-specific transcripts from the two replicates were then clustered and the local alignment with the female genome repeated to generate a list of putative female-specific scaffolds. Ninety-seven candidate female SDR scaffolds were identified in the V4 assembly using this approach.

PCR markers were developed for all candidate female SDR scaffolds using Primer 3 [12] and in silico virtual PCR amplifications were carried out using the e-PCR program [13] to identify oligonucleotide pairs that were predicted to amplify a single region of the female genome but not to amplify from the male genome (Table S1G). Each marker was then tested on genomic DNA of between eight and 57 individuals (at least 18 individuals if the scaffold did not carry a gametologue) of known sex to determine whether the candidate scaffolds were genetically linked to the sex locus (Table S1H). PCR reactions were carried out as described above but in a final volume of 10 μL. The presence or absence of single sex-specific bands was resolved by electrophoresis on 1.5% agarose gels.

Approaches used to improve the assembly of the female SDR

Sex-determining regions are extremely difficult to assemble because they exhibit a high density of repeats. Several strategies were used to improve the assembly of the female SDR (Table S1G). These strategies were applied in parallel with the genetic mapping tests (see above) and focused on scaffolds that had been shown to be sex-linked by these tests. Scaffold structure differed significantly between the V3 and V4 assemblies of the female genome and iterative reciprocal Blasts between the two assemblies allowed many of the putative female SDR scaffolds to be manually extended and linked together. Additional evidence for links between scaffolds were obtained by 1) mapping of raw mate-pair sequence data (using Bowtie) [16] looking for matches to male SDR genes that spanned scaffolds and 2) using scaffold-spanning transcripts predicted by Trinity based on the female RNAseq data. The CLC assembly (Table S1E) also confirmed several links between scaffolds (Table S1F). When a transcript linking two scaffolds was predicted based either on RNAseq data or similarity with a male gametologue, the transcript was verified experimentally by reverse transcriptase PCR amplification from cDNA of between four and eight females using oligonucleotides corresponding to the exons at the two ends of the linked scaffolds (Table S1H). RNA was extracted using the Qiagen Mini kit (`http://www.qiagen.com`) as previously described [1]. RT-PCR reactions were performed using the RT-PCR OneStep kit (QIAGEN, Courtaboeuf, France) following the manufacturers specifications except that we added 5 ng of template RNA and the final reaction volume was 10 μL rather than 50 μL. Thermal cycles were performed as follows: A single cycle for reverse transcription at 50°C for 30 min then an initial denaturing / PCR activation step at 95°C for 15 min followed immediately by 35 cycles of denaturation at 94°C for 30 seconds, annealing at 57°C for 30 seconds, extension at 72°C, and a single final extension step at 72°C for 10 minutes. Single bands were resolved by electrophoresis on a 10 cm long 2% agarose gel. The application of these various approaches allowed the total number of female SDR scaffolds to be reduced to 26.

Annotation of SDR scaffolds

Previously annotated scaffolds from the *Ectocarpus* genome project [2] were considerably improved by integrating transcript information derived from the RNAseq analysis carried out as part of this study and using comparisons of male and female gametologue gene models. The updated gene models can be accessed at the OrcAE database (`http://bioinformatics.psb.ugent.be/orcae/overview/Ectsi`).

The female SDR scaffolds were annotated de novo by running the gene prediction program EuGène [17], which incorporated the signal prediction program SpliceMachine [18], using the optimised Markov models and SpliceMachine splice site predictions derived previously for the male genome sequence [2]. Gene prediction incorporated extrinsic information from mapping of the RNAseq data onto the female-specific scaffolds. Both male and female SDR gene models were manually curated using the raw, mapped RNAseq data, Cuffdiff and Trinity transcript predictions and comparisons between the male and female haplotypes.

The female SDR scaffolds were added to the complete male genome scaffolds to produce a "hybrid genome", which served as the basis for further reference-based analyses. Assembly of the mature gametophyte RNAseq data using this hybrid reference genome as a template was carried out using TopHat (v2.0.3) [7, 8] and Cufflinks (v2.0.2) [8, 9]. For TopHat, the maximum value for multihits per read was set at 40. For both TopHat and Cufflinks the maximum intron size was set at 26,000 bp and annotations were used to guide mapping and transcript assembly against the "hybrid genome".

A first reference assembly of the mature gametophyte RNAseq data, using only the male annotations, was performed with this "hybrid genome". Results of this first assembly were used to confirm automatic gene predictions and attempt to identify new genes. New genes specific to the female were annotated and an annotation file was created. A second assembly was then carried out using the "hybrid genome" reference with both the male and the new female annotation files to compute the abundance of female genes (FPKM) with more accuracy.

Pseudogenes were identified manually by comparing SDR sequences with genes in the public databases. An additional screen for pseudogenes was carried out by blasting male protein sequences against the genomic sequence of the female SDR and vice versa. All sequences that had been annotated as "gene" or "TE" were excluded from this latter analysis using Maskseq and RepeatMasker respectively. We defined a pseudogene operationally as a fragment of nucleotide sequence that resembled a protein sequence in the public databases but was truncated due to the presence of stop codons or frameshifts.

Figure S6 provides a schematic overview of the numbers of loci annotated in the male and female SDR haplotypes, together with information about homology relationships between SDR loci and with autosomal genes. The male SDR haplotype contains 22 protein coding genes and one pseudogene, the female 17 protein-coding genes and seven pseudogenes. Three of the pseudogenes are probably remnants of transposable elements, despite being single copy in the genome, because they share homology with typical transposon proteins such as transposases. The putative transposon remnants were excluded from the analysis of synonymous site divergence (Figure 5D) but were included in gene counts and other statistics (Table 1).

Homologous genes present in both the male and female haplotypes of the SDR were considered to be gametologues (i.e. male and female alleles of the same ancestral gene) if they were detected as matches in a reciprocal Blastp search against the SDR scaffolds (E value cutoff: $1.e^{-4}$). The same criterion was used to identify homologues of SDR genes located outside the SDR (Table S2).

Identification of transposons and other repeated sequences in the SDR

An *Ectocarpus*-specific TE-library (described in [2]), compiled with REPET [19], was used to annotate SDR transposons. TEs were further annotated by running the de novo annotation software Repclass [20] with default parameters. The annotation data from REPET and Repclass were then merged using a custom script. The script retained the REPET annotation when there was a conflict. The autosomes, the PAR and the male and female SDR haplotypes were screened for TEs by running RepeatMasker on each of these genomic compartments using the TE library described above. A custom script was used to parse the RepeatMasker output and to count the total sequence length of each TE category in each genomic compartment. This value was divided by the total length of each compartment (excluding Ns) to calculate the percentage of each region that corresponded to TE sequence.

Intra-haplotype sequence similarity

Sequence similarities within the male and female SDR haplotypes were analysed using a custom Perl code [21]. This code used BLAST (`http://blast.wustl.edu`) and a moving window system to compare 5 kbp sequence segments, in steps of 2 kbp, with the rest of the SDR sequence to detect repeated regions within either the male or the female SDR haplotype (i.e. the analysis was carried out separately for the male and female haplotypes). The sliding window Blast analysis was performed on sequences in which the transposable elements had been masked. By default, the threshold for sequence identity was fixed to 97%. When the threshold was reduced to 50%, the same result was obtained.

Global expression of autosomal versus SDR genes

Transcript abundances, calculated using the male and female mature gametophyte RNAseq data and expressed as FPKM, were compared between autosomal genes and male and female SDR genes. Genes with FPKM values that were greater than twice the standard deviation or equal to zero were removed, and male and female data were pooled. The 95% bootstrap intervals for the means of the two groups did not overlap, indicating that the means were significantly different.

Quantitative reverse transcriptase PCR analysis of SDR gene transcript abundances during the Ectocarpus life cycle

The abundance of male and female SDR gene transcripts during the *Ectocarpus* life cycle was assessed by RT-QPCR. Primer pairs were designed to amplify regions of the 3'UTR or the most 3' exon of the gene to be analysed (Table S4C). In silico virtual PCR amplifications were carried out using the e-PCR program [13] and both the male and female genome sequences to check the specificity of oligonucleotide pairs. RT-QPCR analysis was carried out for 13 male SDR genes and 11 female SDR genes (Figures S4). The remaining SDR genes could not be analysed either because they had very small exons, which posed a problem for primer design, or it was not possible to obtain a single amplification product.

For the RT-QPCR analysis, total RNA was extracted using the Plant RNeasy extraction kit (Qiagen, Courtaboeuf, France) from at least three biological replicates for each of four stages of the life cycle: immature gametophyte, mature gametophyte, partheno-sporophyte and diploid heterozygous sporophyte (Figure 1A). The RNA was treated with RNAse-free DNAse-I according to the manufacturer's instructions (Qiagen) to remove any contaminating DNA and stored at -80°C. The concentration and integrity of the RNA was checked using a NanoDrop 2000 spectrophotometer (ThermoScientific) and by agarose gel electrophoresis. A control PCR without reverse transcriptase was performed to ensure absence of contaminating DNA. For each sample, up to 1 μg of RNA was reverse-transcribed to cDNA using oligo-dT and the Superscript II RT kit (Life Technologies, Gaithersburg, MD, USA) according to the manufacturer's instructions and the cDNA was diluted with water to 1.2 ng equivalent RNA.μL^{-1}.

RT-QPCR was carried out using the ABsolute™ QPCR SYBR® Green ROX Mix (ThermoScientific) in a Chromo4™ thermocycler (BioRad Laboratories) and data were analysed with the Opticon monitor 3 software (BioRad Laboratories) or a LightCycler® 480 multiwell plate 384, on a LightCycler® 480 Real- Time PCR System (Roche Diagnostics, Mannheim, Germany), using the LightCycler® 480 SYBR Green Master mix (Roche Diagnostics, Mannheim, Germany).

For each gene, amplification efficiency (always between 80% and 110%) was tested using a dilution series of male or female *Ectocarpus* genomic DNA (15 ng to 0.006 ng), each dilution being tested in duplicate. Using these genomic DNA dilutions, a standard curve was established for each gene, allowing quantification. Amplification specificity was tested with a dissociation curve. The housekeeping gene ELONGATION FACTOR 1α (EF1α) [3] was used to normalise transcript abundance values. The normalized data correspond to means ±S.E. from three to four independent biological replicates, each of which was calculated from three technical replicates.

To test for the difference in gene expression between the different life cycle stages (immature and mature gametophyte, partheno-sporophyte and sporophyte), a one-way analysis of variance (ANOVA) was performed for the 13 male and 11 female SDR genes. All ANOVAs were conducted using the one way ANOVA procedure implemented in MINITAB (version 13.2 MiniTab Inc. 1994, State College USA). Data were log-transformed in order to meet the normality and homoscedasticity requirement of

ANOVA and multiple comparisons of means were performed using the Fisher method (for the gene FeV4Scaf35_1, as log-transformed data did not meet the homoscedasticity requirement, the Mood non-parametric was performed).

In order to compare the global level of expression of SDR genes in diploid (sporophyte) and in haploid stages (gametophytes and partheno-sporophyte), expression values for each SDR gene were normalized by their mean expression across all stages. To meet the normality and homoscedasticity, male data were boxcox transformed. Normality was tested using the D'Agostino and Pearson omnibus test and homoscedasticity using the Fisher-Snedecor test (F-test). The difference in transcript abundance between diploid and haploid stages was tested using the t-test for the male data and a Welch modified t-test for the female data. All statistical tests were performed using the GraphPad Prism software (`http://www.graphpad.com`). The difference between haploid and diploid stages was significant in male ($p<0.0001$) and female ($P<0.0015$) samples.

The differences in gene expression between sexes (male and female), between ploidy levels (SP and mGa) and their interactions were also tested by a two-way ANOVA using the Glm procedure of MINITAB (version 13.2 MiniTab Inc. 1994, State College USA). Data were boxcox transformed in order to meet the normality and homoscedasticity requirement of ANOVA. No significant difference was detected between the two sexes ($p=0.11$), but transcripts of SDR genes were significantly more abundant in haploid mGA than in diploid SP ($P< 10^{-4}$). This difference was mainly due to increased SDR gene transcript abundance in male gametophytes.

Construction of phylogenetic trees for an SDR and an autosomal gene

Coding sequences from a single exon of two genes making the gametologue pair (Esi68_0003 (male) and FeV4scaf15_1 (female)) were amplified in three *Ectocarpus* lineages and a related brown alga Sphaerotrichia divaricata (C. Agardh) Kylin. The *Ectocarpus* strains described as lineages probably represent separate species based on sequence divergence of autosomal genes, morphology and on sexual crossing experiments. We retain the term lineage since only three *Ectocarpus* species currently have species status [22]. Two of the recognised species, *Ectocarpus siliculosus* (lineage 1a) and *Ectocarpus croaniorum* (lineage 2b) were included in the analysis, and we also used the sequenced strain Ec32 that belongs to lineage 1c. The autosomal tree was constructed using sequences amplified from the same samples using ITS2 nuclear DNA primers. The strains and lineages used are described in Table S1 and the primers are described in Table S3. The PCR conditions were as described above in the "Identification and mapping of the male and female SDR haplotypes" section, but with a final volume of 20 μL to allow for DNA sequencing. Single bands were directly sequenced in the forward and reverse direction. Sequences were edited using Codon Code sequence aligner and the evolutionary history was inferred using both the neighbour-joining method and PhyML in MEGA5 [23] to ensure that both approaches resulted in the same topology. The resulting phylogenetic trees are shown in Figure 5B and 5C.

Synonymous divergence

To estimate synonymous divergence rates the coding sequences of gametologue pairs were translated to protein sequence, and an alignment was performed in the program Seaview [24] using the default Muscle parameters. Alignments were further confirmed using Prank [25] for verification. Those alignments with large regions that were poorly aligned were further analysed using Gblocks [26] in order to eliminate these sections. The aligned protein sequences were then back-translated to coding sequence and synonymous divergence (dS) between gametologous gene pairs was calculated using Codeml in PAML

version 4 [27]. The resulting values were plotted against the gene coordinates the male V chromosome in order to visualise the spread of the data according to gene position and detect whether their positions were organised in relation to the degree of divergence from their female counterpart. Analyses of the data using Kmeans in the program Rv3.0.1 was able to resolve both K of one and two. A Mann-Whitney U test of the resolution K2 was not significant therefore the null hypothesis that the data are formed from a single population was retained. This implies an absence of gene strata on the male V chromosome.

To search for potential gene conversion events dS values were calculated on an exon-by-exon basis (Figure 3). Potential gene conversion events could then be identified by looking for marked variations in dS along the length of each gene.

Estimating the age of the Ectocarpus SDR

The complete nuclear coding sequences of *Phaeodactylum tricornutum*, *Thalassiosira pseudonana* and *Ectocarpus siliculosus* (autosomes) were obtained from the Hogenom database version 6 [28], together with coding sequence data from a further 63 Stramenopile species (following those used in Silberfeld *et al.* [29] and Brown and Sorhannus [30]), which were downloaded from Genbank (`http://www.ncbi.nlm.nih.gov/genbank/`). Clusters of potentially orthologous genes were identified using the program Silix [31] and deduced protein sequences derived from the above sequences. Alignments of the potential orthologues were then created using Muscle and the topology of the resulting sequence clusters were reconstructed using Phyml under a GTR model with an estimated value for gamma and with five different classes [32]. These clusters were then analysed using the program TPMS [33], which can identify the sub-clusters with potential orthologues and hence facilitates the elimination of paralogues. Clusters that contained at least one gene from *Ectocarpus* and one gene from each of the two diatom species, *Phaeodactylum* and *Thalassiosira*, were retained for further analysis.

A blastx search of the *Ectocarpus* genes retained in the clusters was performed against RNA-Seq and Sanger EST data from nine phaeophyceae species and significant matches (E-value $1e^{-10}$) were added to the cluster data. The phylogenies of the gene clusters were further verified manually and those that did not correspond to the species phylogeny were eliminated. In total, 54 clusters of genes from *Ectocarpus*, the two diatoms and 8 phaeophyceae species were obtained, making up 183 pairs. The orthology of the *Ectocarpus* and both diatom sequences was further checked for each of the 54 clusters using the information available in the gene family database Hogenom 6, which includes most the fully sequenced eukaryotic genomes. We searched the phylogenetic trees of the 54 *Ectocarpus* genes in Hogenom 6 and checked whether the diatom sequences that we identified as orthologous were actually the most closely related to the *Ectocarpus* sequence in the those trees. Orthology of the *Ectocarpus* and both diatom sequences was confirmed for 27 clusters, partly confirmed or possible for 10, and not confirmed for the remaining clusters. Only the diatom sequences with confirmed or possible orthology were retained for the subsequent steps of the analysis (i.e. 147 sequence pairs).

A pairwise alignment of the *Ectocarpus* genes with all of the identified orthologous genes from each cluster was then carried out using Prank [25], and alignments were cleaned using Gblocks with highly stringent parameters (maximum number of contiguous non-conserved sites = 2) as we found that less stringent parameters returned less reliable dS values [26, 27]. The programs Codeml and Yn00 from PAML version 4 [27] were then run on each gene pair in order to calculate pairwise dS values. Pairs with aberrant dS-yn00 values (i.e. 99) were excluded, as well as pairs with very high dS-ML values (>20) as these high values probably resulted from convergence problems with divergent sequences in the case of codeml, or to the presence of hidden paralogy among the brown algal sequences. A total of 137 pairs

were available for further analysis. The resulting dS values were plotted against the divergence times (Figure S5) estimated by Silberfeld *et al.* [29] and Brown and Sorhannus [30][30].

Codon usage analysis

Optimal codons were identified by comparing the codon usage of highly expressed genes (ribosomal genes) with the rest of the genome using the multivariate approach described in Charif *et al.* [34] (see also `http://pbil.univ-lyon1.fr/members/lobry/repro/bioinfo04/`). A conservative set of 10 optimal codons was identified, together with two additional, less conservative sets of 14 and 27 optimal codons (Figure S2E andS2F). A custom script was used to estimate the frequency of these optimal codons (Fop) in *Ectocarpus* coding sequences. The Fop values were correlated with RNAseq expression levels (Figure S2E). We retained the set of 27 optimal codons for the analysis of the SDR genes as it gave the highest Spearman rho values. The Fop values in each compartment (autosomes, PAR and SDR) for the three sets of codons are shown in Figure S2F.

Sex-determination in strains carrying different numbers of U and V chromosomes

Diploid gametophytes carrying both the U and the V chromosome can be constructed artificially using the ouroboros (oro) mutant, and these strains are phenotypically male [1]. To determine whether the dominance of the male haplotype was dose dependent, we constructed triploid (UUV) gametophytes, again using oro mutant strains (Figure S1). We isolated 10 independent UV diploid strains (Ec581 to Ec591) and seven independent UUV triploid strains (Ec761 to Ec767) using zygote isolation methods described in Coelho *et al.* [35]. The ploidy of one representative each of the haploid, diploid and triploid *Ectocarpus* strains was verified using a FACSort flow cytometer (`http://www.bsbiosciences.com`; Table S1I). Nuclei were isolated by cutting the filaments with a razor blade and adding nuclei buffer [30 mM MgCl, 120 mM trisodium citrate, 120 mM sorbitol, 55 mM 4-(2-hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES) pH 8, 5 mM EDTA supplemented with 0.1% (v/v) Triton X-100 and 5 mM sodium bisulphite], and DNA content was measured immediately by flow cytometry. Between 600 and 13,200 nuclei were analyzed for each sample. Gametophytes from the male Ec32 strain were used as an internal reference [36, 37]. The nucleic acid-specific stain SYBR Green I (`http://www.invitrogen.com`) was used at a final dilution of 1:10,000. All UV and UUV strains were phenotypically male indicating that the dominance of the male haplotype over the female haplotype is not dose dependent.

HMG-domain genes present in the Ectocarpus genome

A survey for HMG-domain genes identified a total of 13 genes in the *Ectocarpus* genome including the male-specific gene Esi0068_0016, which is the only HMG-domain gene inside the SDR (Table S4A, Figure S4D). Based on the gametophyte RNAseq data, none of the 12 autosomal HMG-domain genes exhibited differential expression in male versus female *Ectocarpus*.

References

1. Coelho SM, Godfroy O, Arun A, Le Corguillé G, Peters AF, et al. (2011) Ouroboros is a master regulator of the gametophyte to sporophyte life cycle transition in the brown alga ectocarpus. Proc Natl Acad Sci U S A 108: 11518–11523.

2. Cock JM, Sterck L, Rouzé P, Scornet D, Allen AE, et al. (2010) The ectocarpus genome and the independent evolution of multicellularity in brown algae. Nature 465: 617–621.

3. Peters AF, Scornet D, Ratin M, Charrier B, Monnier A, et al. (2008) Life-cycle-generation-specific developmental processes are modified in the immediate upright mutant of the brown alga ectocarpus siliculosus. Development 135: 1503–1512.

4. Heesch S, Cho GY, Peters AF, Le Corguillé G, Falentin C, et al. (2010) A sequence-tagged genetic map for the brown alga ectocarpus siliculosus provides large-scale assembly of the genome sequence. New Phytol 188: 42–51.

5. Coelho SM, Scornet D, Rousvoal S, Peters NT, Dartevelle L, et al. (2012) How to cultivate ectocarpus. Cold Spring Harb Protoc 2012: 258–261.

6. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, et al. (2011) Full-length transcriptome assembly from rna-seq data without a reference genome. Nat Biotechnol 29: 644–652.

7. Trapnell C, Pachter L, Salzberg SL (2009) Tophat: discovering splice junctions with rna-seq. Bioinformatics 25: 1105–1111.

8. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, et al. (2012) Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. Nat Protoc 7: 562–578.

9. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, et al. (2010) Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28: 511–515.

10. Anders S, Huber W (2010) Differential expression analysis for sequence count data. Genome Biol 11: R106.

11. Dittami SM, Proux C, Rousvoal S, Peters AF, Cock JM, et al. (2011) Microarray estimation of genomic inter-strain variability in the genus ectocarpus (phaeophyceae). BMC Mol Biol 12: 2.

12. Rozen S, Skaletsky H (2000) Primer3 on the www for general users and for biologist programmers. Methods Mol Biol 132: 365–386.

13. Schuler GD (1997) Sequence mapping by electronic pcr. Genome Res 7: 541–550.

14. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, et al. (1987) Mapmaker: an interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. Genomics 1: 174–181.

15. Le Bail A, Dittami SM, de Franco PO, Rousvoal S, Cock MJ, et al. (2008) Normalisation genes for expression analyses in the brown alga model ectocarpus siliculosus. BMC Mol Biol 9: 75.

16. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short dna sequences to the human genome. Genome Biol 10: R25.

17. Foissac S, Gouzy J, Rombauts S, Mathé C, Amselem J, et al. (2008) Genome annotation in plants and fungi: Eugene as a model platform. Current Bioinformatics 3: 87–97.

18. Degroeve S, Saeys Y, De Baets B, Rouzé P, Van de Peer Y (2005) Splicemachine: predicting splice sites from high-dimensional local context representations. Bioinformatics 21: 1332–1338.

19. Flutre T, Duprat E, Feuillet C, Quesneville H (2011) Considering transposable element diversification in de novo annotation approaches. PLoS One 6: e16526.

20. Feschotte C, Keswani U, Ranganathan N, Guibotsy ML, Levine D (2009) Exploring repetitive dna landscapes using repclass, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biol Evol 1: 205–220.

21. Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.

22. Peters AF, Van Wijk SJ, Cho GY, Scornet D, Hanyuda T, et al. (2010) Reinstatement of ectocarpus crouaniorum thuret in le jolis as a third common species of ectocarpus (ectocarpales, phaeophyceae) in western europe, and its phenology at roscoff, brittany. Phycological research 58: 157–170.

23. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.

24. Gouy M, Guindon S, Gascuel O (2010) Seaview version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27: 221–224.

25. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci U S A 102: 10557–10562.

26. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17: 540–552.

27. Yang Z (2007) Paml 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.

28. Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, et al. (2009) Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10 Suppl 6: S3.

29. Silberfeld T, Leigh JW, Verbruggen H, Cruaud C, de Reviers B, et al. (2010) A multi-locus time-calibrated phylogeny of the brown algae (heterokonta, ochrophyta, phaeophyceae): Investigating the evolutionary nature of the "brown algal crown radiation". Mol Phylogenet Evol 56: 659–674.

30. Brown JW, Sorhannus U (2010) A molecular genetic timescale for the diversification of autotrophic stramenopiles (ochrophyta): substantive underestimation of putative fossil ages. PLoS One 5.

31. Miele V, Penel S, Duret L (2011) Ultra-fast sequence clustering from similarity networks with silix. BMC Bioinformatics 12: 116.

32. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.

33. Bigot T, Daubin V, Lassalle F, Perrière G (2013) Tpms: a set of utilities for querying collections of gene trees. BMC Bioinformatics 14: 109.

34. Charif D, Thioulouse J, Lobry JR, Perrière G (2005) Online synonymous codon usage analyses with the ade4 and seqinr packages. Bioinformatics 21: 545–547.

35. Coelho SM, Scornet D, Rousvoal S, Peters N, Dartevelle L, et al. (2012) Genetic crosses between ectocarpus strains. Cold Spring Harb Protoc 2012: 262–265.

36. Bothwell JH, Marie D, Peters AF, Cock JM, Coelho SM (2010) Role of endoreduplication and apomeiosis during parthenogenetic reproduction in the model brown alga ectocarpus. New Phytol 188: 111–121.

37. Peters AF, Marie D, Scornet D, Kloareg B, Mark Cock J (2004) Proposal of ectocarpus siliculosus (ectocarpales, phaeophyceae) as a model organism for brown algal genetics and genomics. Journal of Phycology 40: 1079–1088.

38. Stache-Crain B, Müller DG, Goff LJ (1997) Molecular systematics of ectocarpus and kuckuckia (ectocarpales, phaeophyceae) inferred from phylogenetic analysis of nuclear-and plastid-encoded dna sequences. Journal of Phycology 33: 152–168.

Supplementary figures

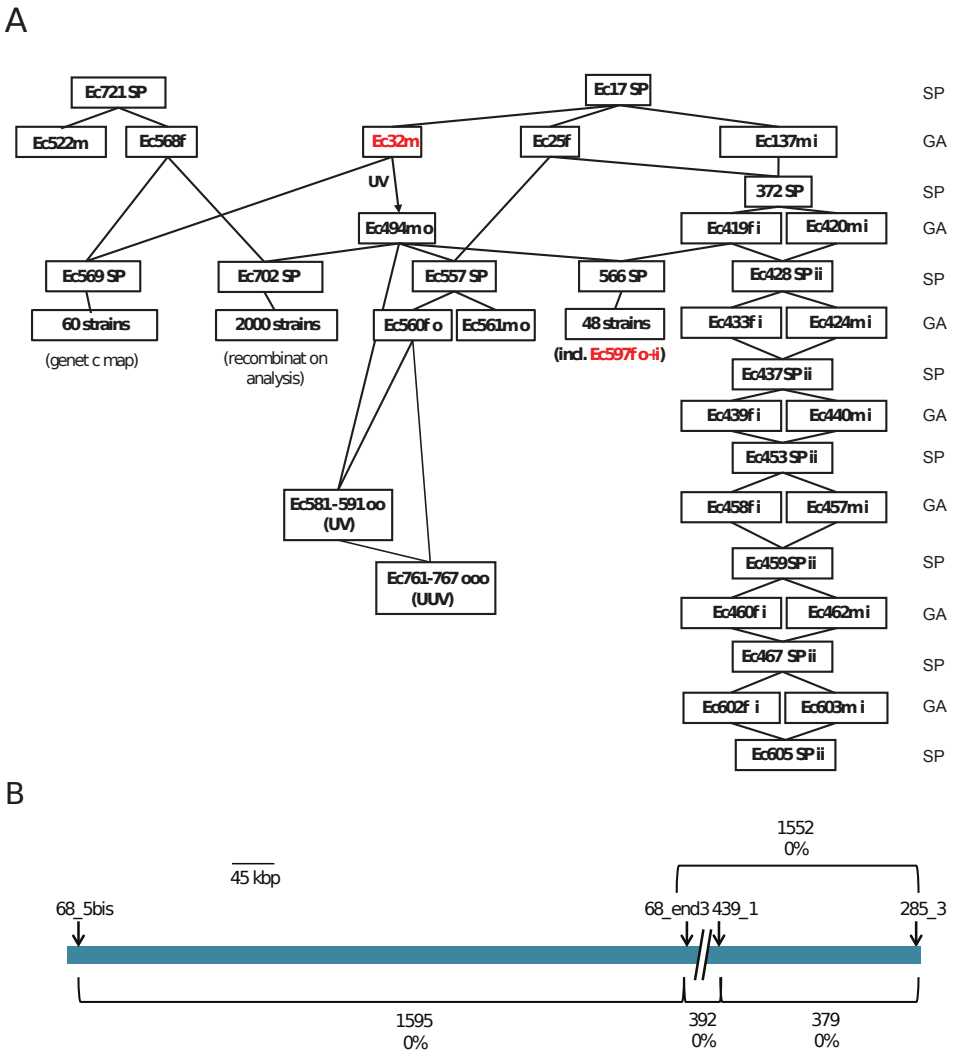


Figure S1 *Ectocarpus* strain pedigree and SDR recombination analysis. Related to Figure 1. **(A)** Pedigree of the *Ectocarpus* strains used in this study. SP, diploid sporophyte; m, male; f, female; i, *immediate upright* mutant; ii, homozygous *immediate upright* diploid; o, *ouroboros* mutant; oo, ooo, diploid and triploid homozygous *ouroboros* mutants; (U), presence of one U sex chromosome; (V), presence of one V sex chromosome. The genomes of strains indicated in red have been sequenced. **(B)** Diagram indicating the extent of recombination between markers located inside the SDR. The number of individuals used to assay for recombination between each pair of markers and the percentage of recombination detected are indicated. See Table S3 for the coordinate position of each marker on its respective scaffold.

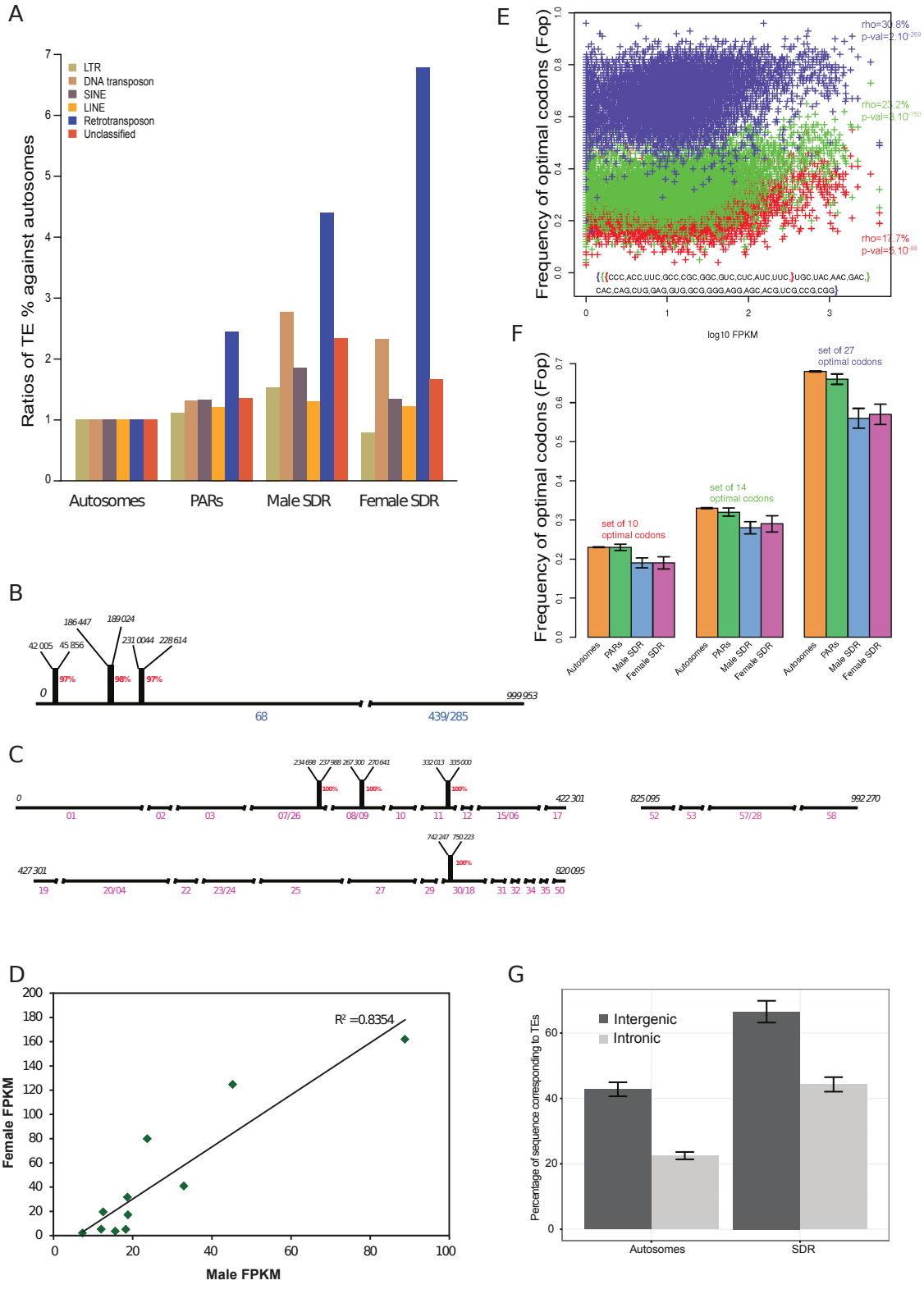

A
LTR
DNA transposon
SINE
LINE
Retrotransposon
Unclassified
Ratios of TE % against autosomes
Autosomes
PARs
Male SDR
Female SDR
B
186 447
189 024
42 005
45 856
231 0044
228 614
97%
98%
97%
0
999 953
68
439/285
C
234 698
237 988
267 300
270 641
332 013
335 000
100%
0
422 301
825 095
992 270
01
02
03
07/26
08/09
10
11
12
15/06
17
52
53
57/28
58
742 247
750 223
427 301
820 095
19
20/04
22
23/24
25
27
29
30/18
31
32
34
35
50
D
R² =0.8354
Female FPKM
Male FPKM
E
Frequency of optimal codons (Fop)
rho=30.8%
p-val=2.10-269
rho=23.2%
p-val=3.10-150
rho=17.7%
p-val=5.10-88
{{{CCC,ACC,UUC,GCC,CGC,GGC,GUC,CUC,AUC,UUC,}UGC,UAC,AAC,GAC,}
CAC,CAG,CUG,GAG,GUG,GCG,GGG,AGG,AGC,ACG,UCG,CCG,CGG}
log10 FPKM
F
Frequency of optimal codons (Fop)
set of 10
optimal codons
set of 14
optimal codons
set of 27
optimal codons
Autosomes
PARs
Male SDR
Female SDR
G
Intergenic
Intronic
Percentage of sequence corresponding to TEs
Autosomes
SDR

Figure S2 Analysis of *Ectocarpus* SDR, PAR and autosomal genes. Related to Figure 2. **(A)** Over-representation of DNA corresponding to different transposable element classes in sex chromosome domains compared to autosomes. Histogram showing percent of transposable element DNA per kilobase for six different classes of transposable element expressed as a ratio with respect to the value calculated for the autosomes. The data used were the same as for Figure 2C: cumulative length of DNA corresponding to each class of transposable element in each genomic compartment divided by the length (excluding Ns) of the compartment. Values greater than one indicate an increased abundance of the corresponding class of transposon in the sex chromosome domain compared to the autosomes. **(B)** Intra-haplotype sequence similarity within the female SDR. Black vertical bars indicate regions that are similar (100%). The scaffold coordinates are indicated in italics and the scaffold number in pink. Note that the order of the female scaffolds is arbitrary. **(C)** Intra-haplotype sequence similarity within the male SDR. Black vertical bars indicate regions that are similar. The supercontig coordinates are indicated in italics and the supercontig number in blue. Note that the orientation of sctg_68 with respect to sctg_439and285 is arbitrary. **(D)** Correlation between the expression of male and female gametologues (FPKM). Linear regression is shown (R^2=0.8354). **(E)** Codon usage in *Ectocarpus*. 3 sets of optimal codons were obtained from a multivariate analysis: a conservative set of 10 optimal codons (red), a permissive set of 27 optimal codons (blue), and an intermediate set of 14 optimal codons (green), see Supplementary notes for details. Correlation between each gene's Frequency of Optimal Codons (Fop) and its log10-transformed expression (in FPKM) in mature male gametophytes. Spearman's rho values for each set of optimal codons and their corresponding p-values are indicated on the right. The optimal codons are indicated. **(F)** Median Fop in coding regions of autosomes, PAR, male and female SDR haplotypes for each set of optimal codons. Error bars indicate 95% confidence intervals around medians. **(G)** Percent of TE sequence in intronic and intergenic regions in *Ectocarpus* autosomes and in the male SDR haplotype (bootstrap r=1000).

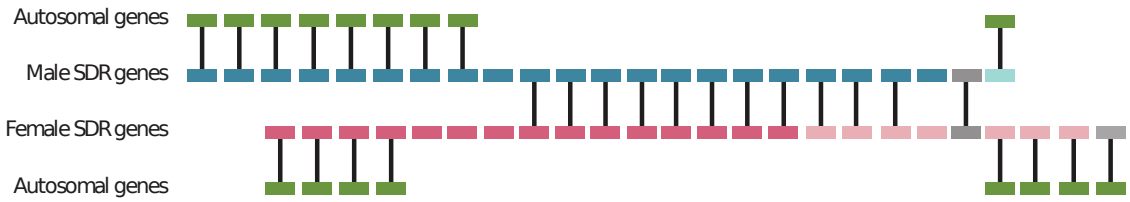


Figure S3 Schematic diagram showing homology relationships between male and female SDR genes and autosomal genes. Related with Figure 3 and Figure 1. Autosomal genes are shown in green, male and female SDR genes are shown in pink and blue respectively, with putative functional genes in dark blue or dark pink and pseudogenes in light blue or light pink. Putative transposon remnants are shown in grey. A green box indicates the existence of at least one homologue in the PAR or on an autosome (i.e. outside the SDR). The existence of an autosomal or PAR homologue is indicated only for genes that do not have a gametologue. All homology relationships were defined as corresponding to a Blastp e-value of less than 10^{-4} when sequences were blasted against the complete set of *Ectocarpus* predicted proteins. Note that the order of the genes is not intended to correspond to their locations in the genome.

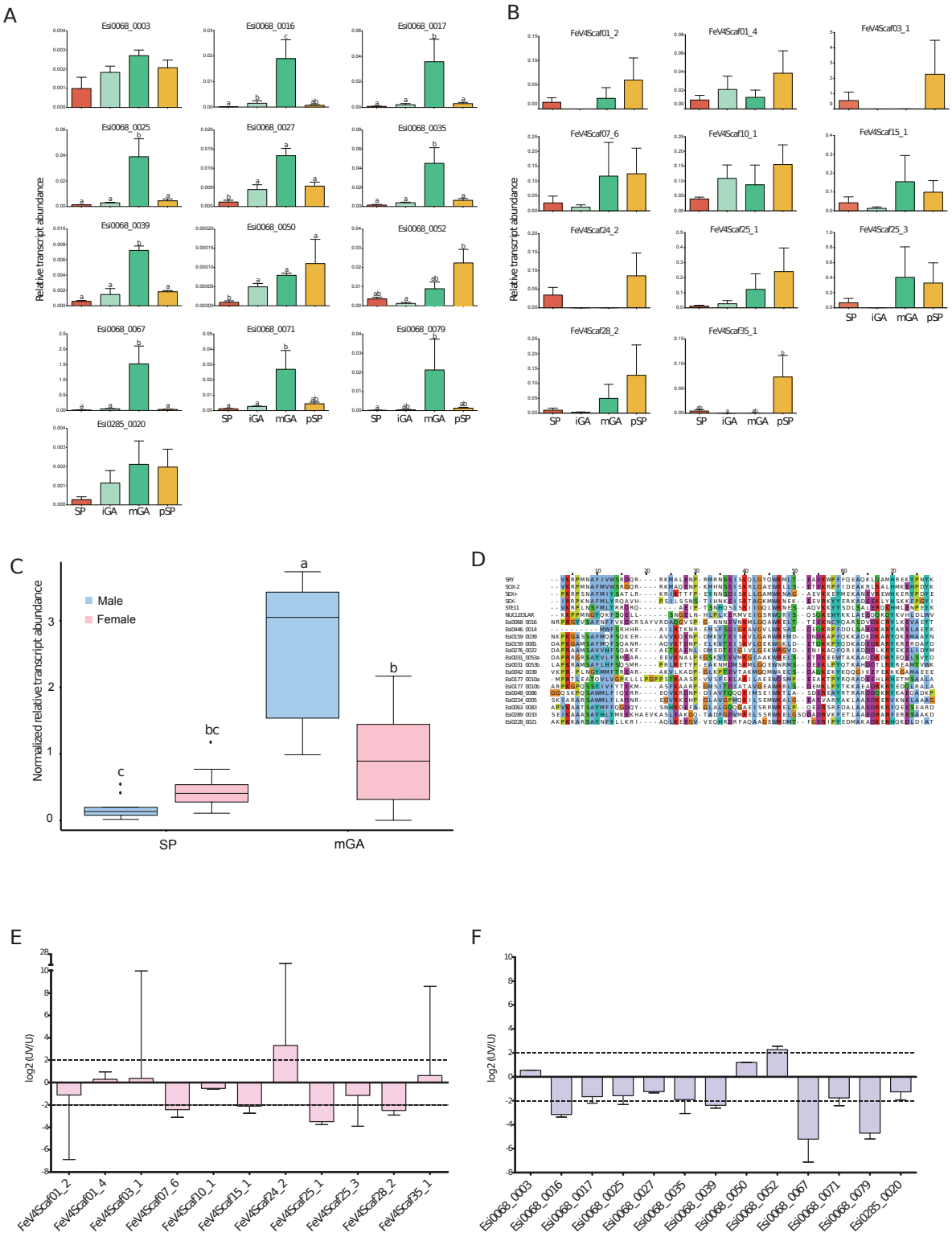

A
Esi0068_0003
Esi0068_0016
Esi0068_0017
Esi0068_0025
Esi0068_0027
Esi0068_0035
Esi0068_0039
Esi0068_0050
Esi0068_0052
Esi0068_0067
Esi0068_0071
Esi0068_0079
Esi0285_0020
Relative transcript abundance
SP
iGA
mGA
pSP
B
FeV4Scaf01_2
FeV4Scaf01_4
FeV4Scaf03_1
FeV4Scaf07_6
FeV4Scaf10_1
FeV4Scaf15_1
FeV4Scaf24_2
FeV4Scaf25_1
FeV4Scaf25_3
FeV4Scaf28_2
FeV4Scaf35_1
C
Male
Female
Normalized relative transcript abundance
SP
mGA
D
E
log2 (UV/U)
F

Figure S4 SDR gene expression during the life cycle. Related to Figure 4. **(A)** Female SDR gene expression during the life cycle of *Ectocarpus*. SP, diploid heterozygous sporophyte; iGA, immature gametophyte; mGA, mature gametophyte; pSP, partheno-sporophyte. Bars with different letters indicate statistical significance (P<0.05). **(B)** Male SDR gene expression during the life cycle of *Ectocarpus*. SP, diploid heterozygous sporophyte; iGA, immature gametophyte; mGA, mature gametophyte; pSP, partheno-sporophyte. Bars with different letters are statistically different (P<0.05). **(C)** Differences in male and female SDR gene expression in the diploid sporophyte (SP) and haploid mature gametophyte (mGA). Identical letters above the error bars indicate that mean values are not significantly different. **(D)** HMG domain alignment. Multiple alignment of the HMG domains of all the HMG domain proteins encoded by the *Ectocarpus* genome together with HMG domains from human and fungal proteins. The human sequences are Sex-determining region Y (HsSRY, CAA37790.1), Sex-determining region Y- box2 (HsSOX2, CAA83435.1) and Nucleolar transcription factor 1 (HsNTF1, 1608205A). The fungal sequences are *Phycomyces blakesleeanus* Sex+ (PbSex+, ABX27912.1) and Sex- (PbSex-, ABX27909.1), and *Schizosaccharomcyes pombe* Ste11 (SpSTE11, CAA77507.1). The *Ectocarpus* Esi0068_0016 gene is located in the SDR. The *Ectocarpus* proteins Esi0032_0053 and Esi0177_0010 are predicted to have two homeodomains (a and b). **(E)** Ratio of the abundance of female SDR gene transcripts in diploid gametophytes (genotypically UV and phenotypically male) compared with mature haploid female gametophytes (genotypically U and phenotypically female). Transcript abundance was measured by quantitative RT-PCR. **(F)** Ratio of the abundance of male SDR gene transcripts in diploid gametophytes (genotypically UV and phenotypically male) compared with mature haploid male gametophytes (genotypically V and phenotypically male). Transcript abundance was measured by quantitative RT-PCR.

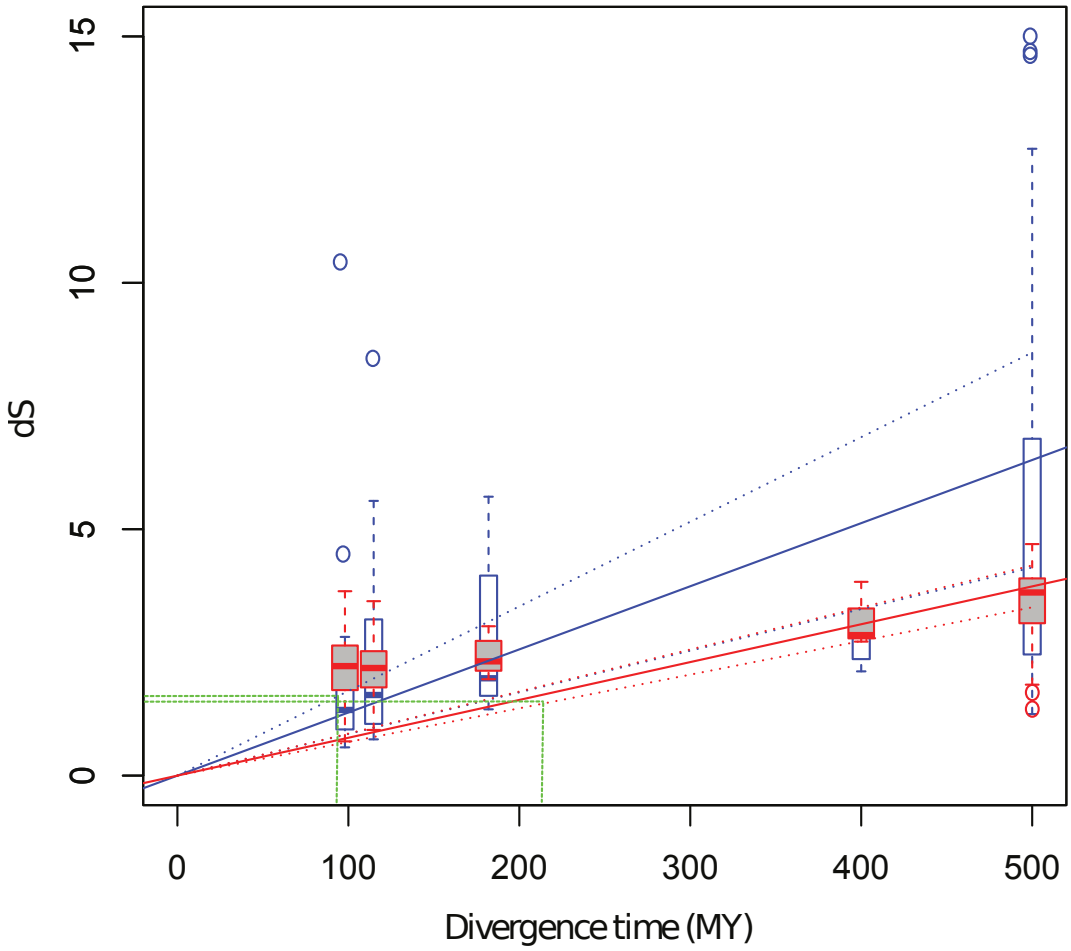


Figure S5 Estimation of the age of the *Ectocarpus* SDR. Related to Figure 5 Box and whisker plot showing calculated dS values for *Ectocarpus*-Stramenopile pairs against literature divergence times (from Silberfeld et al., 2010; Brown & Sorhannus, 2010). Blue boxes show dS values calculated using codeml, red boxes show dS values calculated using yn00. The respective linear regressions through the origin are shown as solid lines, with 99% confidence intervals drawn in dotted lines. The green dashed lines show the lowest and highest 99% confidence ages from Codeml and yn00, representing an estimate of the age of *Ectocarpus* SDR at around 95-215 My. The ages where the codeml and yn00 estimates overlap are approximately 170-190 My. Please note that for clarity the y-axis is limited to 15 (two data points are not represented).

Supplementary tables

Table S1 Ectocarpus strains, sequencing and mapping of the SDRs

Table S1A *Ectocarpus* strains used in this study

Strain	Details	Use in this study	Sex	Reference
Ec32	Male genome sequenced strain	Male reference strain (genome sequence, qPCR)	male	[2]
Ec597	Female genome sequenced strain	Female reference strain (genome sequence, qPCR)	female	this study
Ec603	Male inbred line (8 generations)	RNA-seq and sexual dimorphism analyses	male	this study
Ec602	Female inbred line (8 generations)	RNA-seq and sexual dimorphism analyses	female	this study
Ec87	Sister of Ec32	RT-QPCR	female	this study
Progeny of Ec569	60 strains used for the genetic map	Genetic mapping of the SDR	n/a	[4]
Progeny of Ec702	2000 strains segregating population produced from the heterozygous strain Ec702	Recombination analysis	n/a	this study
Ec568	Outcrossing line	Comparative genome hybridisation	female	[4]
Ec 581-591	Diploid homozygous *oro* mutant gametophytes	Dominance of male haplotype over female	Genotypically UV, phenotypically male	[1]

Table S1A (continued)

Ec761-767	Triploid homozygous *oro* mutant gametophyte (oro,oro,oro)	Dominance of male haplotype over female	Genotypically UUV; phenotypically male	this study
Rb1	Field collected gametophyte	1a strain (Esil 1a_male) used in Figure 3	Male	this study
Da1	Field collected gametophyte	1a strain (Esil 1a_female) used in Figure 3	Female	this study
Has08-5-3	Field collected gametophyte	2b (Ecrou_male) strain used in Figure 3	Male	this study
Ec499	Field collected gametophyte	2b (Ecrou_female) strain used in Figure 3	Female	this study
Sdiv_male	Field collected gametophyte	Male *Sphaerotrichia divaricata* (C. Agardh) Kylin strain used in Figure 3	Male	this study
Sdiv_female	Field collected gametophyte	Female *Sphaerotrichia divaricata* (C. Agardh) Kylin strain used in Figure 3	Female	this study

Table S1B PCR-based markers used to map male SDR scaffolds

N°	Marker	Scaffold	Coordinates on scaffold	Primer 1	Primer 1 sequence	Primer 2	Primer 2 sequence
1	M_68_3	sctg_68	351402	M_68_3F	GTGAGAGAAACAACAGAGCAATACAG	M_68_3R	ATGGAACCGCAGACAACAAGC
2	sctg68_end3	sctg_68	677743	sctg68_end3F	TGGCAACAACACCACCTTTA	sctg68_end3R	AGGACCAGTGAACCATCCTG
3	M_68_4	sctg_68	651823	M_68_4F	AAACACCTCCCAACCAACCAATC	M_68_4R	AACGCAACGAGCAACCTTCC
4	sctg68_5bis	sctg_68	15181	sctg68_5bisF	GGCGCGAAGAAATATATGGA	sctg68_5bisR	CGAGAGTACTGGCCTTTTCG
5	M68_27ex4	Sctg_68	214090	M68_27ex4F	TGCTTGCAACTGGTGACTTC	M68_27ex4R	CGATCATGATGGCAACAACT
6	M68_35ex2	sctg_68	253503	M68_35ex2F	CGTGAACCAGCTTATGGTCA	M68_35ex2R	TCGTTGCAACATCCCAGTAA
7	M_285_1	sctg_285 and439	145553	M_285_1F	CAAGCCAACCCACAACTTTT	M_285_1R	GTTGTCATTCGGCGTAGGAT
8	M_439_1	sctg_285 and439	95398	M_439_F	CACCACCGTCACGTATGAAG	M_439_R	AGCCCTTGTAGACCCAAGGT

Table S1C Evidence used to link male scaffolds

N°	Scaffold 1	Scaffold 1 orientation	Scaffold 2	Scaffold 2 orientation	Scaffolds linked by a miniBAC
1	sctg_439	antisense	sctg_285	sense	KY0AFIP A75YN02

Gametologue	Scaffold spanning gene	Scaffold spanning cDNA detected by RT-PCR	RT-PCR primer 1	RT-PCR primer 1 sequence	RT-PCR primer 2	RT-PCR primer 2 sequence
FeV4scaf04_1	Esi0285_0001	yes	439-285aF	CGATG GCGAA ATAAA AGTGG	439-285aR	AGGTT GGAAA TTGTG CTTGG

Table S1D List of strains of known phenotypic sex (determined by crosses with reference strains) that were used to verify sex linkage of male and female SDR scaffolds

Strain	Species	Lineage	Phenotypic sex	Genotypic sex	Marker used
Ecsil Nap EA1 f	Esil	1a	f	f	68_27ex4, 68_35ex2, 07238, Fe6_15ex5
Ec sil Nap D-A2 f	Esil	1a	f	f	68_27ex4, 68_35ex2, 07238, Fe6_15ex5
Ecsil Na 108 f	Esil	1a	f	f	68_27ex4, 68_35ex2, 07238, Fe6_15ex5
Ec597	Esp	1c	f	f	68_27ex4, 68_35ex2, 07238
Ec25	Esp	1c	f	f	68_27ex4, 68_35ex2, 07238
Ec568	Esp	1c	f	f	68_27ex4, 68_35ex2, 07238
Ec87	Esp	1c	f	f	68_27ex4, 68_35ex2, 07238
ec 467-U15-2	Esp	1c	f	f	68-4, 439_01, scaffold14696
ec 467-u13-5	Esp	1c	f	f	68-4, 439_01, scaffold14696
ec 467-u14-2	Esp	1c	f	f	68-4 , 439_01, scaffold14696
ec 467-u19-5	Esp	1c	f	f	68-4 , 439_01, scaffold14696
EcPH11 85	Ecro	2c	f	f	68_27ex4, 68_35ex2, 07238
EcPH11 113	Ecro	2c	f	f	68_27ex4, 68_35ex2, 07238
Ec32	Esp	1c	m	m	68_27ex4, 68_35ex2, 07238
ec 467-u18-4	Esp	1c	m	m	68-4 , 439_01, scaffold14696
ec 467-u9-1	Esp	1c	m	m	68-4 , 439_01, scaffold14696

Table S1D (continued)

ec 467-u16-5	Esp	1c	m	m	68-4, 439_01, scaffold14696
EcPH11 106	Ecro	2c	m	m	68_27ex4, 68_35ex2, 07238
EcPH11 112	Ecro	2c	m	m	68_27ex4, 68_35ex2, 07238
Tam18b	Esp	2d	m	m	68_35ex2; 07238
Bft15b	Esp	2d	m	m	68_35ex2, 07238
temp19c	Esp	2d	m	m	68_35ex2; 07238
Tam12b	Esp	2d	m	m	68_35ex2;07238
tam2b	Esp	2d	m	m	68_35ex2; 07238
Ec sil Nap R-B1 m	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
Ecsil Na 70 m	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
Ecsil Na 166 m	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
Ec sil Nap EA2 m	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
PH11-138-1	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
PH11-138-6	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
PH11-133-1	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
PH11-s#2A-38U1-1	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
PH11-s#2A-38U1-7	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238
PH11-s#2A-38U1-3	Esil	1a	m	m	68_27ex4, 68_35ex2, 07238

Esil, *Ectocarpus siliculosus*; Ecro, *Ectocarpus crouanorium*; Esp, *Ectocarpus sp.* (no species name attributed); f, female; m, male. Strain lineages are based on Stache-Crain *et al.* [38].

Table S1E Assembly statistics for the genome sequence of the female *Ectocarpus* strain Ec597

		V3 assembly	V4 assembly	CLC assembly
	Assembler	Velvet	Velvet + Gap-Closer	CLC
	Sequence data	454 + paired ends + one mate pair library	454 + paired ends + both mate pair libraries	Paired end data + 454
Contigs (>200 bp)	Number	67,301	33,071	98,137
	Avg size (bp)	2,373	4,648	2,025
	N50 size (bp)	3,706	7,075	5,052
	Cumulative size (bp)	159,686,990	153,699,260	198,775,585
Scaffolds (>200 bp)	Number	44,110	18,835	n/a
	Avg size (bp)	5,402	12,107	n/a
	N50 size (bp)	16,323	25,114	n/a
	Cumulative size (bp)	238,264,795	228,041,014	n/a

Data is provided for the intermediate V3 assembly, the final V4 assembly and an assembly carried out with the CLC assembler using only 454 and paired-end Illumina data.

Table S1F Evidence used to link female scaffolds

N°	Scaffold 1	Scaffold 1 orientation	Scaffold 2	Scaffold 2 orientation	Scaffolds linked by the CLC assembly	Scaffold spanning gametologue match	Scaffold spanning gene	Scaffold spanning cDNA detected by RT-PCR	RT-PCR primer 1	RT-PCR primer 1 sequence	RT-PCR primer 2	RT-PCR primer 2 sequence
1	FeV4 Scaf10	antisense	FeV4 Scaf14	sense	yes	Esi0068_0035	FeV4 scaf10_1	yes	FeV4 Scaf10-04aF	GCAGTTCTGGAAAACGGAAA	FeV4 Scaf10-04aR	GTAGGCGCTTGGTGAAGAGA
2	FeV4 Scaf15	antisense	FeV4 Scaf6	sense	yes	Esi0068_0003	FeV4 scaf15_1	yes	FeV4 Scaf15-06aF	AACTGGCGAAGTGGACGA	FeV4 Scaf15-06aR	ATACCTGTCCGGCTCCTTG
3	FeV4 Scaf7	sense	FeV4 Scaf26	sense		Esi0068_0025	FeV4 scaf07_6	yes	FeV4 Scaf07-26aF	GCGACATACGCACTGTTTTG	FeV4 Scaf07-26aR	GGGCCTATCCTCCTCAATGT
4	FeV4 Scaf23	sense	FeV4 Scaf24	sense			FeV4 scaf23_1	yes	FeV4 Scaf23-24fF	CTCCCAACGGTCTTGTTCTC	FeV4 Scaf23-24fR	GAAAAGTGCATGGTCCCAAC

Table S1F (continued)

5	FeV4 Scaf57	sense	FeV4 Scaf28	sense	Esi0285 _0010	FeV4 scaf28_2	yes	FeV4 Scaf28 -57fF	CAATC GGAGA TCGTG CTGT	FeV4 Scaf28 -57fR	GCTTC ACGTC CACCT CCTC
6	FeV4 Scaf8	antisense	FeV4 Scaf9	antisense	Esi0285 _0020	FeV4 scaf08_1	yes	FeV4 Scaf09 -08aF	GCCTG GAAGG AGTGA AGGA	FeV4 Scaf09 -08aR	ACCTG GGATC AATGT TGTCT G
7	FeV4 Scaf30	antisense	FeV4 Scaf18	sense		FeV4 scaf30_1	yes	FeV4 Scaf18 -30fF	CTGAG TGGAA CTGAC GTGGA	FeV4 Scaf18- 30fR	CGACC CTCCT TGATA CGTTG
8	FeV4 Scaf20	sense	FeV4 Scaf4	sense		FeV4 scaf04_1	yes	FeV4 Scaf20 -04aF	CCTGG TTGTT GTGCT TGGA	FeV4 Scaf20 -04aR	TGCCA TCTTT TCACG TCTGT

Table S1G List of female SDR scaffolds showing their sizes in base pairs and the number of individuals used to map each scaffold to the sex locus

Scaffold name	length	Carries a gametologue?	Annotated gene on scaffold?	Size of mapping population
FeV4Scaf01	98133	yes	yes	17
FeV4Scaf02	17680	yes	yes	25
FeV4Scaf03	51021	yes	yes	17
FeV4Scaf04and20	78753	yes	yes	23
FeV4Scaf06and15	45363	yes	yes	15
FeV4Scaf07and26	57992	yes	yes	15
FeV4Scaf08and09	41097	yes	yes	8
FeV4Scaf10	14559	yes	yes	8
FeV4Scaf11	25512	no	no	28
FeV4Scaf12	6590	no	no	57
FeV4Scaf17	16002	no	no	55
FeV4Scaf18and30	33261	yes	yes	28
FeV4Scaf19	19637	no	yes	57
FeV4Scaf22	16900	yes but fragment	yes	28
FeV4Scaf24and23	37411	no	yes	55
FeV4Scaf25	63333	yes	yes	18
FeV4Scaf27	52265	no	yes	55
FeV4Scaf28and57	66706	yes	yes	16
FeV4Scaf29	11350	no	no	55
FeV4Scaf31	11604	no	no	55
FeV4Scaf34	2875	no	no	55
FeV4Scaf35	3359	no	yes	18
FeV4Scaf50	3460	no	no	55
FeV4Scaf52	24537	no	yes	55
FeV4Scaf53	18312	no	no	55
FeV4Scaf58	42520	no	yes	55
Total	**860232**			

Table S1H PCR-based markers used to map female SDR scaffolds

N°	Marker	Scaffold	Primer 1	Primer 1 sequence	Primer 2	Primer 2 sequence
1	scaffold15594	FeV4Scaf01	scaffold15594F	AGACGCGAAG AACGAACACT	scaffold15594R	CCGCGATTTT GTGCTCGTAG
2	scaffold15594a	FeV4Scaf01	scaffold15594aF	TTCGCTTTGA TTGGGCTATG	scaffold15594aR	ACCAAGTTTC TGGGCAAGGT
3	scaffold5692Da	FeV4Scaf01	scaffold5692DaF	GCTCGGCTTG ATAGGTCATG	scaffold5692DaR	AGGTTATTGG CCTTGGTTGC
4	Ecfemscaf72511	FeV4Scaf01	Ecfemsca72511R	GGTCGGAGAG CGTAAGAGGT	Ecfemsca72511L	TAGGGTTGTT TTGCGATGGA
5	Ecfemscaf4078	FeV4Scaf01	Ecfemsca4078R	TAGGGTTGTT TTGCGATGGA	Ecfemsca4078L	GGTCGGAGAG CGTAAGAGGT
6	Ecfemscaf70269	FeV4Scaf01	Ecfemsca70269R	GTCGTGGTTG TTCGTGTGTG	Ecfemsca70269L	GTGAGCATTT GGCTGGAAGA
7	Ecfemscaf2091	FeV4Scaf01	Ecfemsca2091R	ACCAAGTTTC TGGGCAAGGT	Ecfemsca2091L	TTCGCTTTGA TTGGGCTATG
8	Ecfemscaf1174	FeV4Scaf01	Ecfemsca1174R	ACCAAGTTTC TGGGCAAGGT	Ecfemsca1174L	TTCGCTTTGA TTGGGCTATG
9	Ecfemscaf82697	FeV4Scaf01	Ecfemsca82697R	GGGTCGTTCT TTCTGTGCTG	Ecfemsca82697L	TTCAGTTTTC ATGCCGTTCC
10	06636	FeV4Scaf01	06636F1	GACGCAACAG GGAGGCACCA ATA	06636R1	TGCGCGTAAC AGGGGAAAA ACAA
11	scaffold4647	FeV4Scaf02	scaffold4647F	GAATCGGGCT CACGAGAGAG	scaffold4647R	ACGAATTGAT TAAGCGGCGC
12	scaffold4647a	FeV4Scaf02	scaffold4647aF	CAGGTGGGTT GTCATGTGTG	scaffold4647aR	TACCTACGCC CGAATGAATG
13	scaffold14727	FeV4Scaf02	scaffold14727F	GGCGCGGTGA AATACGTTAC	scaffold14727R	GAGGATCGGC AAAATCGCAC

Table S1H (continued)

14	scaffold14727a	FeV4Scaf02	scaffold14727aF	ACGGTAGGGT CGGAATCAAG	scaffold14727aR	TTGCATGTGT GCGAGTCTGT
15	Ecfemscaf61128	FeV4Scaf02	Ecfemsca61128R	GGCAGACCAC AACAGGGTAG	Ecfemsca61128L	TAAGCAAGGCT CAACCAGGA
16	07238	FeV4Scaf02	07238F1	AAGGAACGCA AACCGCCGAA ATA	07238R1	CTCCATCCCC AACGTTGTCT CTGTG
17	scaffold18795a	FeV4Scaf03	scaffold18795aF	TTGGCGAAAC GAAATCAAAG	scaffold18795aR	TGGTGTAATC GTCCCTGCTC
18	Ecfemsca1372	FeV4Scaf04and20	Ecfemsca13724R	ATTTCTGGTG AAAGCGCAAA	Ecfemsca13724L	CACGAAGGAG GGGGTAAAAA
19	Ecfemsca7690	FeV4Scaf04and20	Ecfemsca7690R	TGCGACGAGA AAGAAGGAAA	Ecfemsca7690L	AATTGAAACC CCGTCCAATC
20	Ecfemsca5503	FeV4Scaf04and20	Ecfemsca5503R	TGCGACGAGA AAGAAGGAAA	Ecfemsca5503L	AATTGAAACC CCGTCCAATC
21	Ecfemsca30656	FeV4Scaf04and20	Ecfemsca30656R	TCTTCCAGAC GGTGGAGTTG	Ecfemsca30656L	TGGTTACGGC AGCTTCATTT
22	scaffold14581	FeV4Scaf04and20	scaffold14581F	CTGTGATTGT TGCGCACACA	scaffold14581R	CGTGTGAGTG GTTTTGGCTG
23	scaffold14581b	FeV4Scaf04and20	scaffold14581bF	CGGTACTCCC TCACCACTCA	scaffold14581bR	GGGCAAAAGA GCAACACAAA
24	scaffold10110a	FeV4Scaf06and15	scaffold10110aF	TTTGTTGTTG GACCCCTTTG	scaffold10110aR	CCTTCGTTTT CCTCCTCTGG
25	scaffold4518	FeV4Scaf06and15	scaffold4518F	GATCCGTTGT TGGCGTATGC	scaffold4518R	GTGCAACGTG CCTGGAATTT
26	scaffold16822a	FeV4Scaf06and15	scaffold16822aF	GCCAACACAG CACTACACGA	scaffold16822aR	GGGAATAAAC ACGACCAGCA
27	scaffold4518a	FeV4Scaf06and15	scaffold4518aF	TTGCGGTACT TGTTGCTGTG	scaffold4518aR	AGCGGGAACT AAGCGGTAGA

Table S1H (continued)

28	Ecfemscaf15535	FeV4Scaf06and15	Ecfemsca15535R	TTGTTTCATC GGCAAAAACC	Ecfemsca15535L	CAGGAACCCC CACTGTATGA
29	scaffold7067	FeV4Scaf07and26	scaffold7067F	GAAACGAGCA ACGATCGACG	scaffold7067R	TCCACTTTCA CTGACGACGG
30	Ecfemsca109773	FeV4Scaf07and26	Ecfemsca109773R	GTGCGATACA CCGAACGAAC	Ecfemsca109773L	CTCCAATCCC CCACTCATTT
31	scaffold18979	FeV4Scaf07and26	scaffold18979F	AAACGACCGC AAGAAACACG	scaffold18979R	GATTCGGCGC ACATTCAGTC
32	Ecfemsca3352	FeV4Scaf08and09	Ecfemsca3352R	CTCTGACGAG CCACATCCTG	Ecfemsca3352L	GTTCTGGACA ACGGTGGAAC
33	Ecfemsca20136	FeV4Scaf08and09	Ecfemsca2013R	CGCACGTAGC TCTTTCGATG	Ecfemsca2013L	AAGGTTGTGC TAGGGGGAGA
34	Ecfemsca23188	FeV4Scaf08and09	Ecfemsca23188R	CGCATTCCGA TTCTCTCCTC	Ecfemsca23188L	CTGCATTCCT CACTCGTTCC
35	Ecfemsca65052	FeV4Scaf10	Ecfemsca5052R	GCCTTCCGTG TGCTAGTCTG	Ecfemsca5052L	GCGTGGGTAG ATGCAGTAGG
36	Ecfemscaf8896	FeV4Scaf11	Ecfemsca8896R	CGCGACCCCC TATCTACTTC	Ecfemsca8896L	AACGCTTCGG AGACTTCACA
37	Ecfemscaf95028	FeV4Scaf11	Ecfemsca95028R	GGTTCGGTCT CTCCTGTTCC	Ecfemsca95028L	CCTCTATTGG CGGACCATCT
38	Ecfemscaf26954	FeV4Scaf11	Ecfemsca26954R	CGCGACCCCC TATCTACTTC	Ecfemsca26954L	AACGCTTCGG AGACTTCACA
39	scaffold2938	FeV4Scaf12	scaffold2938F	CGTTCGTGAC GCAATCGTAC	scaffold2938R	CATCCATCCG ACGGAAGAGG
40	scaffold2938a	FeV4Scaf12	scaffold2938aF	GCGCTGATTG GAAGTGAAAA	scaffold2938aR	CGCAACAACA CAAAGAGCAG
41	scaffold12304	FeV4Scaf17	scaffold12304F	CTATCCTTCC CGCCTCGAAC	scaffold12304R	GCGAACCTGC GTTGTCTTTT

Table S1H (continued)

42	scaffold12304a	FeV4Scaf17	scaffold12304aF	AATTTCAGCT CGCCAAGACA	scaffold12304aR	TCTCCCGTTC GGCTATTTTT
43	scaffold19158	FeV4Scaf18and30	scaffold19158F	TCCGACCAAG TCCTCGTTTG	scaffold19158R	CGCCAGCGAT TCTAACAACG
44	scaffold13317	FeV4Scaf18and30	scaffold13317F	GATGCTCGTT CGTTCGTTCG	scaffold13317R	CCATACCCGC ATCCTCAGAC
45	scaffold13317a	FeV4Scaf18and30	scaffold13317aF	TTACAACAGC CCACCTCACC	scaffold13317aR	GACTGGCGTA CCGAAAACAA
46	scaffold1585a	FeV4Scaf18and30	scaffold1585aF	TGTTGCTTGT GCGACTGTTC	scaffold1585aR	GTCTCTTGGT GATCGCTTGC
47	scaffold14014a	FeV4Scaf19	scaffold14014aF	TGTAAGCGAA GGGAGCAAGA	scaffold14014aR	AGTCTAAGGG CCGGAAACAG
48	scaffold14775a	FeV4Scaf22	scaffold14775aF	GCGGTTGAAA GGAAGAGGAG	scaffold14775aR	AAAAAGGCGA AATGGAGGAA
49	scaffold14775	FeV4Scaf22	scaffold14775F	CTCGCTCCCG CTTATGTGAT	scaffold14775R	TGTCGAAACG CTTGCTGTTG
50	scaffold7638	FeV4Scaf22	scaffold7638F	GCGAAAACCG ACGAAACACA	scaffold7638R	CATTCGGTGT TTCGATCCGC
51	scaffold14696	FeV4Scaf24and23	scaffold14696F	GTGCGGCGAA AATAATCCCC	scaffold14696R	TTGCTTTCTC TGTTGCACGC
52	scaffold6658	FeV4Scaf24and23	scaffold6658F	AAACGAGCCG AGAGATAGCG	scaffold6658R	CCACGAGCTT TGTGTTGGTG
53	scaffold17544e	FeV4Scaf24and23	scaffold17544eF	ACCGCAAAAC AAAAGTGGAA	scaffold17544eR	GCATTCAGGA GCAAGGAGTG
54	scaffold6658a	FeV4Scaf24and23	scaffold6658aF	TCACCAACAC AAAGCTCGTG	scaffold6658aR	CGTGTAACGG CTGCATTTTT
55	scaffold7164	FeV4Scaf25	scaffold7164F	TATGACGCGT GGCCATAGAC	scaffold7164R	CTGTTCGTTC TCCCCGTGAA

Table S1H (continued)

56	scaffold13849	FeV4Scaf25	scaffold13849F	GCGAAGATCG AGATCCGGTT	scaffold13849R	GCGTTCGCAG TAACATCGAC
57	scaffold13849a	FeV4Scaf25	scaffold13849aF	AACGACGGTA GCTTGGGTTT	scaffold13849aR	TCGGTCGGTT TGTGATTTTC
58	scaffold16354	FeV4Scaf27	scaffold16354F	CGCGCAGACC TAGTCATCAT	scaffold16354R	CGCGCGTCCA CTTAAGAATG
59	scaffold10828	FeV4Scaf28and57	scaffold10828F	AAGAGGACGA TGGCTATGCG	scaffold10828R	GAAGCGTTCA TCGGCGTAAC
60	scaffold17248a	FeV4Scaf28and57	scaffold17248aF	CACTCCGTGA ACTTGAACCA	scaffold17248aR	TCGAAGGAGG AAGCAACAAC
61	scaffold17248b	FeV4Scaf28and57	scaffold17248bF	CTTTGTGCTT GGGTTGGATT	scaffold17248bR	ATGTCCGCTC TGTCTTTTCC
62	scaffold16559	FeV4Scaf29	scaffold16559F	GCCTTTCGCG ATGACATCAC	scaffold16559R	GTGAGACGGC CATTCACTGA
63	scaffold14997	FeV4Scaf31	scaffold14997F	CATCGATTCA ACCTGCAGCG	scaffold14997R	CCACACTACG GACGATAGCC
64	scaffold7781	FeV4Scaf34	scaffold7781F	GCCGGCACAT CTACCTAGAC	scaffold7781R	TCGAACCGGT GGTCTTTCTG
65	scaffold4943	FeV4Scaf35	scaffold4943F	CGCATACTTT GTCGAGTGCG	scaffold4943R	AATCGGAGCC ACATCAACGT
66	scaffold1020	FeV4Scaf50	scaffold1020F	TTTTGAAAGC GTTCTGCCCG	scaffold1020R	AAACACGATG GGGCTTTTGC
67	scaffold7381	FeV4Scaf52	scaffold7381F	GCGAAAACCG ACGAAACACA	scaffold7381R	CATTCGGTGT TTCGATCCGC
68	scaffold8468	FeV4Scaf53	scaffold8468F	CTTACTTGCG CATGTCGTCG	scaffold8468R	AGGCCGTGAA CCACATTAGG
69	scaffold18586	FeV4Scaf58	scaffold18586F	TTCTCGGGAC AGAAACCACG	scaffold18586R	GCAGCACGAT GCACATACAG

Table S1I Ploidy of *Ectocarpus* strains determined by flow cytometry

***Ectocarpus* strain**	**Average fluorescence intensity (arbitrary units)**	**Number of cells analysed**	**Ploidy**
Ec32	57	3,111	n
Ec560	52	1,037	n
Ec581	127	868	2n
Ec761	163	1,066	3n

Ec32, haploid wild type gametophytes; Ec560, haploid *oro* mutant strain; Ec581, homozygous, diploid *oro* mutant strain; Ec761, triploid *oro* mutant strain.

Table S2 Male and female SDR genes

Gene	Main ORF (bp)	n. copies in male genome (Blastp)[1]	FPKM male	FPKM female	total n. transcripts (Cuffdiff)	Gametologue/homologue	Function	Gene/Pseudogene/ Non-coding or repeat	upregulated during fertility? (RT-QPCR)	Comment
Esi0068_0003	921	6	18.718	0.000	3	FeV4scaf15_1	GTPase activating protein	Gene	yes	Full-length GTPase coding sequence with conserved domain.
Esi0068_0016	942	2	11.507	0.000	2	Esi0159_0039*	High mobility group protein	Gene	yes	Not conserved enough to determine if coding region is complete.
Esi0068_0017	1566	7	8.246	0.000	5	Esi0214_0031*	Conserved hypothetical protein	Gene	yes	Not conserved enough to determine if coding region is complete. No domains found.
Esi0068_0025	2181	4	18.841	0.000	1	FeV4scaf07_6	Nucleotide-diphospho-sugar transferase domain protein	Gene	yes	Not conserved enough to determine if coding region is complete.
Esi0068_0027	1056	1	18.314	0.000	2	FeV4scaf25_1	MEMO-like domain protein	Gene	yes	Full-length MEMO protein coding sequence with conserved domain.
Esi0068_0035	795	6	45.297	0.016	1	FeV4scaf10_1	Chloroplast clp protease P	Gene	yes	Full-length clp protease coding sequence with conserved domain.
Esi0068_0039	3033	1	15.650	0.000	3	FeV4scaf30_1	Conserved hypothetical protein	Gene	yes	Probably full-length coding region. Esi0068_0039 only weakly matches FeV4scaf30_1.
Esi0068_0050	1113	6	0.348	0.000	1	Esi0003_0246	Conserved hypothetical protein	Gene	no	Not conserved enough to determine if coding region is complete. No domains found.
Esi0068_0052	804	15	2.789	0.016	3	Esi0328_0030	Conserved hypothetical protein	Gene	no	Full-length protein coding sequence. No domains found.
Esi0068_0058	1314	>250	32.940	0	2	FeV4scaf01_4	STE20-like serine/threonine kinase	Gene	no	Full-length STE-20 protein coding sequence with conserved domain.
Esi0068_0061	141	1	32.940	0	1	no hit	Hypothetical protein	Gene	-	Small transcription unit within the transcribed region of Esi0068_0058 and sharing one of the Esi0068_0058 UTR exons.
Esi0068_0062	264	1	0.331	0	2	no hit	Hypothetical protein	Gene	-	Small monoexonic gene.

Esi0068_0064	540	2	0.111	0	1	Esi0242_0044 (LG30 PAR)*	Putative thioesterase/thiol ester dehydrase-isomerase	Gene	-	Full-length protein coding sequence.
Esi0068_0067	1110	2	88.803	0	2	FeV4scaf02_1	Homoaconitate hydratase	Gene	yes	Full-length protein coding sequence.
Esi0068_0068**	609	3	0.114	0	1	FeV4scaf25_3	Conserved hypothetical protein	Transposon remnant	-	Similar to Topoisomerase III. Putative transposon.
Esi0068_0071	1110	2	20.987	0	2	Esi0099_0001*	Ubiquitin C-terminal hydrolase	Gene	yes	Full-length protein coding sequence.
Esi0068_0079	2124	62	7.361	0	4	FeV4scaf03_1	Casein kinase	Gene	yes	Coding region is twice as long as casein kinase sequences in the databases.
Esi0439_0009	288	4	2.240	0	3	Esi0119_0058*	Conserved hypothetical protein	Gene	-	Probably full-length coding region. No domains found.
Esi0439_0005	198	2	0.051	0	1	Esi0026_0059*	Related to HSP70	Pseudogene	-	Matches HSP70s but much shorter. Out of frame ORF with protein matches in the Ec32 genome at the 3rd exon.
Esi0285_0001	1299	118	12.084	0	1	FeV4scaf04_1	LRR and regulator of G protein signaling domain protein	Gene	-	Spans sctg_285 and sctg_439. Not conserved enough to determine if coding region is complete. No domains found.
Esi0285_0010	855	39	12.597	0	2	FeV4scaf28_2	Protein phosphatase 2C	Gene	-	Full-length protein coding sequence.
Esi0285_0020	486	13	23.722	0	3	FeV4scaf08_1	RING-type Zinc finger domain	Gene	no	Probably full-length coding region.
Esi0285_0022	1842	7	0	0	1	Esi0119_0021*	Conserved hypothetical protein	Gene	-	Not conserved enough to determine if coding region is complete. No domains found.
FeV4scaf15_1	2547	11	0	31.602	2	Esi0068_0003	GTPase activating protein	Gene	no	Large (500 aa) N-terminal extension compared to database matches.
FeV4scaf07_6	2247	4	0	17.172	1	Esi0068_0025	nucleotide-diphospho-sugar transferase domain protein	Gene	no	Not conserved enough to determine if coding region is complete.
FeV4scaf25_1	735	1	0	5.093	1	Esi0068_0027	MEMO-like domain protein	Gene	no	Full-length MEMO protein coding sequence with conserved domain.

FeV4scaf10_1	795	6	0	124.681	2	Esi0068_0035	Chloroplast clp protease P	Gene	no	Full-length clp protease coding sequence with conserved domain.
FeV4scaf30_1	2730	0	0	3.58259	2	Esi0068_0039	Conserved hypothetical protein	Gene	-	Identical copy of second exon on V4scaffold13516. Not conserved enough to determine if coding region is complete. No domains found. FeV4scaf30_1 only weakly matches Esi0068_0039.
FeV4scaf01_4	1263	>250	0	40.851	2	Esi0068_0058	STE20 protein kinase	Gene	no	Full-length STE-20 protein coding sequence with conserved domain.
FeV4scaf02_1	1219	2	0	162.028	2	Esi0068_0067	Homoaconitate hydratase	Gene	-	Full-length protein coding sequence.
FeV4scaf25_3**	672	2	-	-	1	Esi0068_0068	Conserved hypothetical protein	Transposon remnant	no	Protein is similar size to male gametologue. No domains found.
FeV4scaf03_1	1329	>50	0	2.074	3	Esi0068_0079	Casein kinase	Pseudogene	no	Truncated casein kinase gene (end corresponds to FeV4scaf22_1).
FeV4scaf22_1	195	1	0	8.628	3	no hit	Casein kinase fragment	Pseudogene	-	End fragment of a casein kinase gene. Weak similarity with Esi0068_0079.
FeV4scaf04_1	1113	>50	0	5.210	4	Esi0285_0001	LRR protein	Gene	no	Not conserved enough to determine if coding region is complete.
FeV4scaf28_2	850	25	0	19.520	3	Esi0285_0010	Protein phosphatase 2C	Pseudogene	no	Probably full-length coding region.
FeV4scaf08_1	204	5	0	59.842	2	Esi0285_0020	Fragment of RING-type Zinc finger gene	Pseudogene	-	Shorter than the male gametologue.
FeV4scaf01_2	600	1	-	-	1	Esi0347_0020*	Putative acetyl-coenzyme A transporter	Pseudogene	no	Protein is truncated compared with acetyl-coenzyme A transporters in Genbank.
FeV4scaf03_2	516	6	0.091	0.092	1	Esi0259_0016*	Conserved hypothetical protein	Gene	-	Not conserved enough to determine if coding region is complete. No domains found.
FeV4scaf04_2	423	3	-	-	1	Esi0096_0027 (LG30 PAR)*	Histidine triad protein	Gene	-	Not conserved enough to determine if coding region is complete. Histidine triad domain (IPR).
FeV4scaf19_1	159	0	0	20.466	1	no hit	Hypothetical protein	Gene	-	Not conserved enough to determine if coding region is complete.
FeV4scaf24_2	2271	5	0.000	7.594	2	no hit	Patched domain protein	Gene	no	Not conserved enough to determine if coding region is complete.

FeV4scaf25_2	1053	4	-	-	1	Esi0079_0093 (LG27)*	MULE transposase domain protein	Transposon remnant	-	Contains MULE transposase domain. Protein is truncated compared to Esi0079_0093.
FeV4scaf27_1	395	0	0	1.853	1	no hit	Hypothetical protein	Gene	-	Not conserved enough to determine if coding region is complete.
FeV4scaf28_1	453	1	-	-	1	Esi0313_0028 (LG30 PAR)*	Conserved hypothetical protein	Pseudogene	-	Based on comparison with Esi031-0028, the other half of this gene is on V4.scaffold10906 (equivalent to sctg_313 in the male).
FeV4scaf35_1	663	11	0	1.059	1	Esi0142_0041 (LG12)*	Heat shock protein 70	Pseudogene	-	Incomplete gene at start of scaffold, also appears to be truncated at the C-terminal end.
FeV4scaf52_1	396	7	-	-	1	Esi0194_0022 (LG19)*	Conserved hypothetical protein	Gene	-	Incomplete gene at start of scaffold.
FeV4scaf58_1	762	4	0	3.634	1	Esi0034_0115 (LG25)*	Conserved hypothetical protein	Gene	-	Not conserved enough to determine if coding region is complete.

[1]blastp of protein against male predicted proteome, cutoff E:10^{-4}. LG, linkage group; *not sex linked; **The Esi0068_0068/FeV4scaf25_3 pair of genes were predicted to correspond to transposon remnants and were not, therefore, counted among the gametologues.

Table S3 Primers used to amplify the sequences from an SDR and an autosomal gene that were employed to construct the phylogenetic trees shown in Figure 3

Number	Gene	Primer name	Forward sequence	Reverse sequence
1	Esi0068_0003	ExEsi68_3ex5	TCCAGTTTGTGATGGACTCG	TGAATAATGCCAGACACACTCTG
2	FeV4scaf15_1	ExFe15_1ex6	CGTGGTGGACTCATTGACTG	GTGCCAGACATACCCTGTAGAAC
3	Esi0005_0027	ExEsi5_27ex2	GAGTTCATCAACGACGAGCA	CTACCCGTTTCCTTGAACCA

Table S4 Expression analysis of SDR genes

Table S4A Complete list of HMG-domain proteins encoded by the *Ectocarpus* genome

Gene	Functional Description	IPR domains	Pfam domains	FPKM male	FPKM female	Differential expression RNAseq?	Best hit blastn female V4 (scaffold/score/E value)
Esi0031_0053	High mobility group protein	High mobility group, superfamily	HMG (high mobility group) box x2	117.2	86.29	no	scaffold8238 / 46 / 0.002
Esi0042_0039	High mobility group and SAP domain protein	DNA-binding SAP; High mobility group, superfamily	SAP domain; HMG (high mobility group) box	48.17	37.82	no	scaffold17135 / 266 / e^{-69}
Esi0048_0086	High mobility group and histone-like transcription factor domain protein	Transcription factor CBF / NF -Y / archaeal histone; High mobility group, superfamily; Histone-fold	HMG (high mobility group) box; Histone-like transcription factor (CBF / NF-Y) and archaeal histone	6.87	9.91	no	scaffold5380 / 111 / $8e^{-23}$

Table S4A (continued)

Esi0063_0083	High mobility group, SNF2 and SLIDE domain protein	SNF2-related; SANT / Myb domain; Helicase, C-terminal; Homeodomain-like; High mobility group, superfamily; Helicase, superfamily 1 / 2, ATP-binding domain; SLIDE domain	HMG (high mobility group) box; SNF2 family N-terminal domain; Helicase conserved C-terminal domain; SLIDE	17.94	14.03	no	scaffold11914 / 957 / 0.0
Esi0068_0016	High mobility group domain protein	High mobility group, superfamily	HMG (high mobility group) box	11.51	0	n.a	-
Esi0159_0039	High mobility group domain protein	High mobility group, superfamily	HMG (high mobility group) box	184.97	154.83	no	scaffold8835 / 448 / e^{-125}
Esi0159_0081	High mobility group domain protein	Structure-specific recognition protein; High mobility group, superfamily; Domain of unknown function DUF1747; SSRP1 domain	Structure-specific recognition protein (SSRP1); Histone chaperone Rttp106-like; HMG (high mobility group) box	40.90	47.35	no	scaffold3529 / 490 / e^{-136}
Esi0177_0010	High mobility group domain protein	High mobility group, superfamily	HMG (high mobility group) box x2	0.68	0.785	no	scaffold14575 / 583 / e^{-165}

Table S4A (continued)

Esi0224_0005	High mobility group and histone-like transcription factor domain protein	Transcription factor CBF / NF-Y / archaeal histone; High mobility group, superfamily; Histone-fold	HMG (high mobility group) box; Histone-like transcription factor (CBF / NF-Y) and archaeal histone	19.52	18.24	no	scaffold3986 / 472 / e^{-131}
Esi0228_0021	High mobility group and PHD zinc finger domain protein	Zinc finger, RING-type; Zinc finger, PHD-type; High mobility group, superfamily; Zinc finger, FYVE / PHD-type; Zinc finger, RING / FYVE / PHD-type; Zinc finger, PHD-type, conserved site; Zinc finger, PHD-finger	Domain of unknown function (DUF1898); PHD-finger	6.91297	3.78089	no	scaffold8714 / 163 / $7e^{-38}$

Table S4A (continued)

Esi0276_0022	High mobility group domain protein	High mobility group, superfamily	High mobility group, superfamily	14.6604	12.2964	no	scaffold14232 / 2139 / 0.0
Esi0289_0033	High mobility group, SNF2 and SLIDE domain protein	SNF2-related; SANT / Myb domain; Helicase, C-terminal; Homeodomain-like; High mobility group, superfamily; Helicase, superfamily 1 / 2, ATP-binding domain; ATPase, nucleosome remodelling ISWI, HAND domain; SLIDE domain	Domain of unknown function (DUF1898); SNF2 family N-terminal domain; Helicase conserved C-terminal domain; SLIDE; HMG (high mobility group) box	20.546	17.8343	no	scaffold12975 / 476 / e-132
Esi0446_0014	High mobility group domain protein	High mobility group, superfamily	High mobility group, superfamily	8.65888	4.57929	no	scaffold5638 / 424 / e^{-117}

Table S4B Differential gene expression in male and female mature gametophytes, obtained by RNAseq analysis

Sex	Total number of expressed genes	Number of genes showing statistically significant sex-biased expression (DESeq)	% sex-biased genes
Male	14192	354	2.5%
Female	14239	234	1.6%

Table S4C List of primer pairs used for the quantitative RT-PCR analysis of male and female SDR gene expression

Gene ID	Primer 1	Primer 1 sequence	Primer 2	Primer 2 sequence	Amplicon size	Analysis
Esi0068_0003	Esi0068_0003F	GCGATGATGG TTGGTATGGT	Esi0068_0003R	CATACGTTGG CTCGTGTGTT	88	Male RT-QPCR
Esi0068_0016	Esi0068_0016F	CTCCCGGAAA CAAACAATGAA	Esi0068_0016R	GTCTGACCGC GCTTGATAAC	75	Male RT-QPCR
Esi0068_0017	Esi0068_0017F	CGTTCAACCA GGAAGGACA	Esi0068_0017R	CGTCCGAAGC TCTGCACTAT	162	Male RT-QPCR
Esi0068_0025	Esi0068_0025F	GTCCGTATGA ATGGCTGGAT	Esi0068_0025R	TTCCTTCGTG TATCGCTTGTT	128	Male RT-QPCR
Esi0068_0027	Esi0068_0027F	CTCGGACTCT GCCTCGAC	Esi0068_0027R	CAGCAGCACA CACCAACTTC	216	Male RT-QPCR
Esi0068_0035	Esi0068_0035F	ATTGCTGTAG GCCACCAACT	Esi0068_0035R	GTTGCGTCGT GCATGTATTC	152	Male RT-QPCR
Esi0068_0039	Esi0068_0039F	AGTCAGGTCG ACGCACAAG	Esi0068_0039R	GCTCCCAACA GAGGACACC	80	Male RT-QPCR
Esi0068_0050	Esi0068_0050F	CTACTGCCTC CACTACGCTTC	Esi0068_0050R	CTGCTCCAAC ATCCTCCATT	203	Male RT-QPCR
Esi0068_0052	Esi0068_0052F	GCGGACGTGT GTATTGTGTT	Esi0068_0052R	TCCTTGCTCG ATAGGCTCTG	193	Male RT-QPCR

Table S4C (continued)

Esi0068_0067	Esi0068_0067F	AAATGATAGG GTACTGGTGG AGAA	Esi0068_0067R	ATACATTCAC AGAGGTCAAC ACG	105	Male RT-QPCR
Esi0068_0071	Esi0068_0071F	AGTACCGTGG AGTTGTGAAGC	Esi0068_0071R	CCTGTCTTAT GACGCACTCG	103	Male RT-QPCR
Esi0068_0079	Esi0068_0079F	TCGCGGATGC CAGACTAT	Esi0068_0079R	GTGCCGGTGA GAGACCTTC	111	Male RT-QPCR
Esi0285_0020	Esi0285_0020F	TCAGGCAGCA AGACTGAGG	Esi0285_0020R	CTGAAGTCCG AACAATGAAGG	172	Male RT-QPCR
FeV4Scaf01_2	FeV4Scaf01_2F	GCGTGATGGA TGAGTGGAC	FeV4ScaR01_2R	TGCGAAGAAA GTATCGCTTG	187	Female RT-QPCR
FeV4Scaf01_4	FeV4Scaf01_4F	ATTTCGGCTT GGGTGTTG	FeV4ScaR01_4R	GTATGCCTCG CAGTTGGAAG	156	Female RT-QPCR
FeV4Scaf03_1	FeV4Scaf03_1F	TATACGGCTC AAGGCCACTC	FeV4ScaR03_1R	GCTTCTACCG CTCCAACATC	116	Female RT-QPCR
FeV4Scaf07_6	FeV4Scaf07_6F	GAGAGCGCCT GTTGTATTCG	FeV4ScaR07_6R	TTGATCCCAT GAACGAACG	145	Female RT-QPCR

Table S4C (continued)

FeV4Scaf10_1	FeV4Scaf10_1F	GAAGGAAAGG AGCAAATGG	FeV4ScaR10_1R	CGTCGTTCGC CAGATAAAG	128	Female RT-QPCR
FeV4Scaf15_1	FeV4Scaf15_1F	TCGTAGTGCT GACGGAAGAG	FeV4ScaR15_1R	GGAAGAGATG CGCTAACACC	138	Female RT-QPCR
FeV4Scaf24_2	FeV4Scaf24_2F	AATACTGCGG GTTTCATGGTAG	FeV4ScaR24_2R	GGTAGTTTCC GTTGCTCATCC	178	Female RT-QPCR
FeV4Scaf25_1	FeV4Scaf25_1F	CCGAAAAAGT GGGAAAGAGG	FeV4ScaR25_1R	GGGAGGAGAA CTGAACAATCC	107	Female RT-QPCR
FeV4Scaf25_2	FeV4Scaf25_2F	ATGGGATGGG CAGTGTTC	FeV4ScaR25_2R	GTAACGCTGA ACCGCAAGTC	181	Female RT-QPCR
FeV4Scaf25_3	FeV4Scaf25_3F	ATCCCCGGTT TGAAGAGAG	FeV4ScaR25_3R	TCGTTACAGC CGTCATATCG	136	Female RT-QPCR
FeV4Scaf28_2	FeV4Scaf28_2F	CAGTTGCCCT ATCCGATGAT	FeV4ScaR28_2R	CAGGCCGATC CTAGTCATCT	146	Female RT-QPCR
FeV4Scaf35_1	FeV4Scaf35_1F	ACGGAACGCA CATGAACC	FeV4ScaR35_1R	TATGGCGGTG AACTGATCG	164	Female RT-QPCR

Conclusion

Studies of sex chromosome evolution date back to the beginning of the twentiest century, when Mendel's followers studied the X-linked recessive mode of inheritance of human traits. During the early times of sex chromosome research, the study of a few model species have resulted in hypotheses that could not be thoroughly tested until the advances in large-scale sequencing data in the 2000s (e.g. the sequencing of the human X and Y chromosomes [1, 2], and transcriptomic data thanks to next-generation sequencing (NGS)).

My PhD focused on the human XY system, where most of the data on sex chromosomes are available. I studied how the Y chomosome is rescued by the X. I contributed to an ongoing controversy about Ohno's hypothesis and suggested that there is no global doubling of expression on the X, but dosage-sensitive genes seem to have adapted their expression on a gene-by-gene basis. My results, and that of others [3, 4], raise the question of why X-chromosome inactivation (XCI) is global where dosage compensation requirements are local. In a recent review I proposed explanations for XCI evolution other than dosage compensation of the unicopy X-linked genes. I also looked for evidence of X-Y gene conversion during primate evolution. I was able to find regions of X-Y gene conversion with clear-cut boundaries inside several gametologs, which led me to discuss the evolutionary meaning of these gene conversion events. These two studies suggest that the X chromosome may not be as much a help for the Y as has been suggested.

In a second project, I used an alternative model for studies on sex chromosome evolution: the UV sex chromosomes in the brown alga *Ectocarpus siliculosus*, a species

from a eukaryotic supergroup very distantly related to that of animals (>1 billion years of divergence). This study confirmed that an arrest of recombination and a degeneration of the heterogametic / haploid sex chromosome(s) often happen during sex chromosome evolution. These results are also consistent with the hypothesis that evolutionary strata successively form in a sex chromosome system because of the presence of sexually-antagonistic genes, as *Ectocarpus* displays little sexual dimorphism and relatively small sex-specific regions. Finally, this study also raised doubts on the view that the age of the sex-specific region correlates with its size (together with ratite birds studies [5]), as we found it to be very ancient in *Ectocarpus*: 100-200 million years old.

We are currently living a very stimulating era where more and more genomic data are produced everyday, allowing us: 1) to study more profoundly the sex chromosome evolution mechanisms in the well-known model species, and 2) to compare these mechanisms between very different organisms. This accumulating data will allow us to address questions yet unresolved such as: Are the current and past rates of Y degeneration different between species? How fast and why do ampliconic regions change between Y chromosomes? What is the precise location of the pseudo-autosomal boundary between species, and how is its position determined? What are the proportions of XY, ZW and UV systems in the eukaryotic tree? Why do certain organisms keep the same sex chromosomes for long while others have a high turnover?

However, using NGS to get more sex-linked sequences is likely to be a major methodological challenge: Y chromosomes are full of repeated elements and highly heterochromatinized, and are very difficult to assemble with short NGS reads. Consistent with this idea, D. Page's lab is working on a more accessible version of their BAC approach to sequence Y chromosomes (SHIMS, see [6]), and if the announced third-generation single molecule sequencing technology finally comes out, fascinating discoveries will be made in the field of sex chromosome evolution.

Bibliography

[1] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, et al. (2003) The male-specific region of the human y chromosome is a mosaic of discrete sequence classes. Nature 423: 825–837.

[2] Ross MT, Grafham DV, Coffey AJ, Scherer S, McLay K, et al. (2005) The dna sequence of the human x chromosome. Nature 434: 325–337.

[3] Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, et al. (2012) Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. PLoS Biol 10: e1001328.

[4] Lin F, Xing K, Zhang J, He X (2012) Expression reduction in mammalian x chromosome evolution refutes ohno's hypothesis of dosage compensation. Proc Natl Acad Sci U S A 109: 11752–11757.

[5] Vicoso B, Kaiser VB, Bachtrog D (2013) Sex-biased gene expression at homomorphic sex chromosomes in emus and its implication for sex chromosome evolution. Proc Natl Acad Sci U S A 110: 6453–6458.

[6] Hughes JF, Rozen S (2012) Genomics and genetics of human and primate y chromosomes. Annu Rev Genomics Hum Genet 13: 83–108.

A

ANNEX: The evolution of GC-biased gene conversion in eukaryotes

During my first year of master degree, I made a four months internship supervised by Gabriel Marais and Laurent Duret. I studied the possible presence of GC-biased Gene Conversion (gBGC) along the eukaryotic tree, using 36 completely sequenced genomes. I looked for signatures that gBGC can leave in genomes, but this does not constitute by itself an evidence for active gBGC in these genomes, rather trails for future studies.

This work was completed during my first year of PhD, and Alexandra Popa helped me with the genetic maps analyses. We then wrote the manuscript of our results during my second year of PhD.

After a first submission to *MBE*, this paper was sent to *GBE* for review on the 1st of November 2011, accepted on the 17th of May 2012 and published on the 23rd of May 2012.

Evidence for Widespread GC-biased Gene Conversion in Eukaryotes

Eugénie Pessia[1], Alexandra Popa[1], Sylvain Mousset[1], Clément Rezvoy[1,2], Laurent Duret[1], and Gabriel A. B. Marais[1,3,*]

[1]Université Lyon 1, Centre National de la Recherche Scientifique, UMR5558, Laboratoire de Biométrie et Biologie évolutive, Villeurbanne, Cedex, France

[2]École Normale Supérieure de Lyon, Centre National de la Recherche Scientifique, UMR5668, Laboratoire de l'Informatique du Parallélisme, Lyon, Cedex, France

[3]Present address: Laboratoire Biométrie et Biologie Evolutive (LBBE), CNRS, Université Lyon 1, France

*Corresponding author: E-mail: gabriel.marais@univ-lyon1.fr.

Accepted: May 17, 2012

Abstract

GC-biased gene conversion (gBGC) is a process that tends to increase the GC content of recombining DNA over evolutionary time and is thought to explain the evolution of GC content in mammals and yeasts. Evidence for gBGC outside these two groups is growing but is still limited. Here, we analyzed 36 completely sequenced genomes representing four of the five major groups in eukaryotes (Unikonts, Excavates, Chromalveolates and Plantae). gBGC was investigated by directly comparing GC content and recombination rates in species where recombination data are available, that is, half of them. To study all species of our dataset, we used chromosome size as a proxy for recombination rate and compared it with GC content. Among the 17 species showing a significant relationship between GC content and chromosome size, 15 are consistent with the predictions of the gBGC model. Importantly, the species showing a pattern consistent with gBGC are found in all the four major groups of eukaryotes studied, which suggests that gBGC may be widespread in eukaryotes.

Key words: GC-biased gene conversion, recombination, GC content, chromosome size.

During meiotic recombination, parental chromosomes undergo not only large-scale genetic exchanges by crossover but also small-scale exchanges by gene conversion. These events of gene conversion can be biased. In particular, there is evidence that in some species gene conversion affecting G/C:A/T heterozygous sites yields more frequently to G/C than to A/T alleles, a phenomenon called GC-biased gene conversion (gBGC) (Eyre-Walker 1993; Galtier et al. 2001; Marais 2003; Duret and Galtier 2009a). gBGC is expected to increase the GC content of recombining DNA over evolutionary time and is considered a major contributor to the variation in GC content within and between genomes (Eyre-Walker 1993; Galtier et al. 2001; Marais 2003; Duret and Galtier 2009a). gBGC has caught a lot of attention because it affects the probability of fixation of GC alleles and looks like selection for increasing GC, which can mislead several tests designed to detect positive selection (Galtier and Duret 2007; Berglund et al. 2009; Duret and Galtier 2009b; Galtier et al. 2009; Ratnakumar et al. 2010; Webster and Hurst 2012). It has been demonstrated that gBGC occurs during meiosis in budding yeast (Birdsell 2002; Mancera et al. 2008), and there is strong indirect evidence that this process also affects mammals, where clear-cut relationships between local GC content and recombination rates and many other observations consistent with gBGC have been reported (Galtier 2003; Montoya-Burgos et al. 2003; Spencer et al. 2006; Duret and Arndt 2008; Romiguier et al. 2010). Other studies have investigated gBGC in several organisms such as opossum, chicken, sticklebacks, *Drosophila*, honeybees, *Caenorhabditis elegans*, *Arabidopsis*, wheat, rice, the marine unicellular algae *Ostreococcus*, and the ciliate *Paramecium* (Marais et al. 2001, 2003; International Chicken Genome Sequencing Consortium 2004; Marais et al. 2004; Beye et al. 2006; Galtier et al. 2006; Mikkelsen et al. 2007; Duret et al. 2008; Haudry et al. 2008; Jancek et al. 2008; Escobar et al. 2010; Capra and Pollard 2011; Muyle et al. 2011; Nabholz et al. 2011). However, most of the currently

available data comes from animals and plants, and we lack a global picture on gBGC in eukaryotes.

Here we wanted to investigate whether gBGC has affected genome evolution in other eukaryotic groups. One typical signature of gBGC is that, on the long term, this process leads to a positive correlation between local GC content and recombination rates (reviewed in Marais 2003; Duret and Galtier 2009a). We thus looked for such a relationship in eukaryotic species for which the genome was entirely sequenced. We focused our analyses on taxa for which the genome sequence was assembled and anchored on chromosomes. We included all species available, except for metazoans, which are clearly over-represented in genomic sequence databases, and for which we only selected a representative sample. Our dataset includes 36 species from four of the five major eukaryotic groups: Unikonts, Excavates, Stramenopiles and Plantae ([Keeling et al. 2005], see fig. 1). Recombination data are available for 17 of these species, mostly Metazoan (Unikonts) and Plantae (see table 1). Among these 17 species, 6 show a significant correlation between chromosome-averaged recombination rate and GC content (table 1). Interestingly, out of these six correlations, five are positive. Thus, when a significant correlation is detected, it is in most cases consistent with gBGC. Moreover, the mean correlation coefficient is significantly >0 (0.31, $P = 0.0015$), again consistent with gBGC.

To investigate gBGC in a larger sample of species, including those without recombination data, we used chromosome size as a proxy for recombination rates. It has been shown that chromosome size and recombination rates are inversely correlated in many eukaryotes (e.g., Kaback 1996; Copenhaver et al. 1998; Kaback et al. 1999). This pattern reflects the fact that in many species, the proper segregation of chromosomes during meiosis requires having at least one crossover per chromosome, and that the occurrence of a crossover on a given chromosome decreases the probability of having a second one on the same chromosome (a process termed "crossover interference"). These constraints lead to a lower crossover rate (per Mb) in large chromosomes compared with small ones (Kaback 1996; Copenhaver et al. 1998; Kaback et al. 1999). Among species for which genetic maps are available, we found that in most cases (14/17) chromosome size indeed correlates negatively with recombination rates (table 1), and all significant correlations are negative (7/7). The gBGC model therefore predicts a negative correlation between chromosome size and GC content (although other explanations are possible, see Discussion below). Accordingly, this expected correlation has been found in yeast—for which there is direct evidence of gBGC—and mammals—for which there is strong indirect evidence of gBGC (Bradnam et al. 1999; Meunier and Duret 2004). Table 2 shows that among the 36 eukaryotic species studied, 13 show a significant correlation between chromosome size and chromosome-wide GC content (12 after correction for multiple testing, see table 2). Out of these 13 correlations, 12 are consistent with gBGC—that is, negative. The single exception is *Trypanosoma brucei,* which shows a significant positive correlation between chromosome size and GC content. Figure 2 shows three examples illustrating the different types of situations that we observed: *Leishmania major* (significant negative correlation), *T. brucei* (significant positive correlation) and *Guillardia theta* (no significant correlation).

The evolution of chromosomal GC content can be driven by various processes: point substitutions, deletions, or insertions (including repeated sequences). Interestingly, we observed similar correlations when using GC at third codon position (GC3) instead of total GC content (table 2). Given that third codon positions can only evolve by base replacement, this shows that the observed correlation is due to variation in the pattern of point substitutions, and not to variation in DNA repeat content across chromosomes (table 2). In several cases, the statistical significance of the correlation changed from the total GC content analysis to the GC3 one, but the total number of species showing data consistent with gBGC is similar (significant negative correlation: 13/36, significant positive correlation: 1/36). Both analyses gave qualitatively the same results, with—as expected—changes in statistical significance caused by slight changes of the coefficients of correlation in case of species with low chromosome number (i.e., *Dictyostelium discoideum*, *Sorghum bicolor, T. brucei, Cryptococcus neoformans, Micromonas pusilla)*. *Thalassiosira pseudonana* and *Phaeodactylum tricornutum*, two diatoms with a relatively large number of chromosomes, show results consistent with gBGC only for GC3, which raises the possibility of different mutation patterns affecting coding and noncoding regions in these species.

The fact that about half of the species shows the footprint of gBGC (i.e., a significant negative correlation) may indicate gBGC is absent in the other half. It may also indicate that our approach fails to detect gBGC in many species. Indeed, the statistical significance of the correlations strongly depends on the number of chromosomes. For species with few chromosomes, our ability to detect the signature of gBGC is limited. For instance, *G. theta* shows a strong negative correlation between chromosome size and GC content (fig. 2*c*), but with only three chromosomes, the *P* value is obviously nonsignificant. We thus performed a statistical power analysis using human as a reference (see Materials and Methods). Table 2 shows the statistical power (from 0 to 100%) for all species of our dataset. Most species have too few chromosomes to detect any significant correlation between GC content and chromosome size. Among the 19 species for which the estimated power of our test is $>50\%$, 14 (74%) show a significant correlation with total or third position GC content, and in all cases the correlation is consistent with gBGC. Similarly, another power analysis using a more conservative reference (yeast) revealed that 14 out of the 28 species with a power of $>50\%$ show results consistent with gBGC.

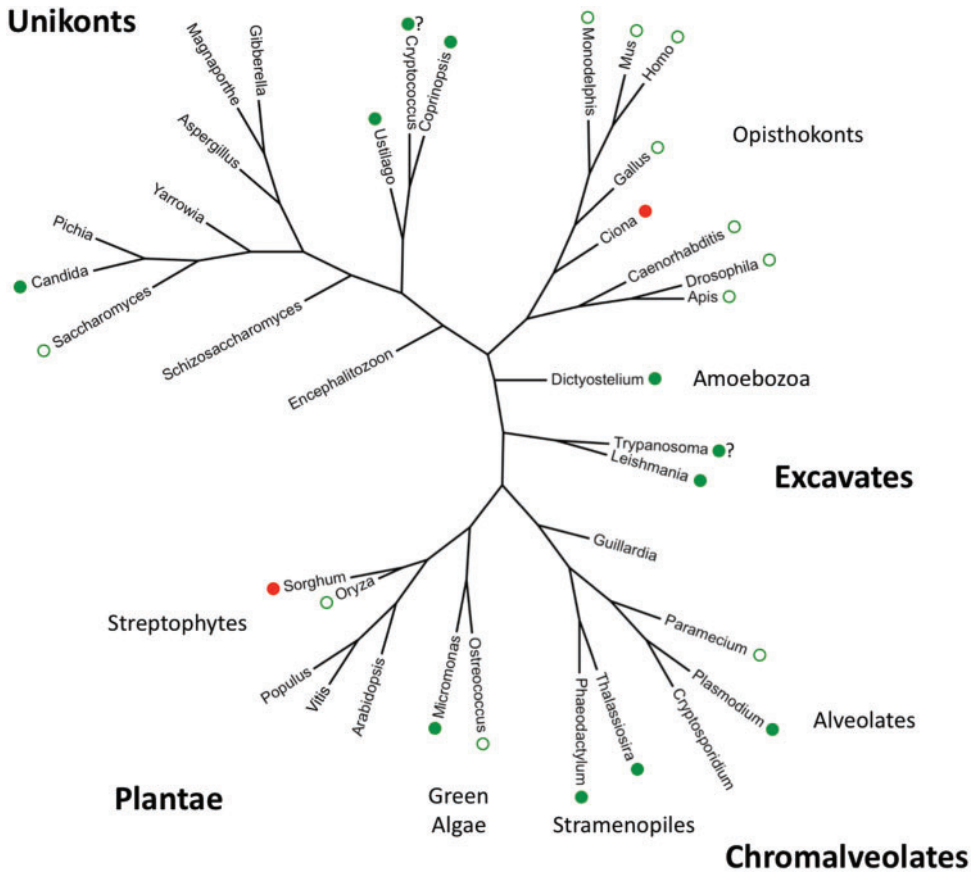


Fig. 1.—Phylogenetic tree of the 36 species studied. Major groups in eukaryotes (see Keeling et al. 2005) are indicated. Green circles indicate significant positive correlations between GC content (total GC content and/or GC3) and recombination rates (measured directly or using chromosome size as a proxy), consistent with gBGC (this work and others). Red circles indicate significant negative correlations between GC content and recombination rates, not consistent with gBGC. Filled circles indicate new observations from the present study. The "?" indicates when results using direct or indirect measures of recombination rates are not fully consistent.

Moreover, the combined analysis of all species indicated a strong significant negative correlation (for total GC content and chromosomes size: P value $= 10^{-50}$, for GC3 and chromosome size: P value $= 10^{-63}$). However, focusing only on the species that show individually nonsignificant correlations, the combined analysis is not significant. There is thus no clear trend emerging from this subset of species.

Given that chromosomal size is only a rough proxy for recombination rate, this result is most likely an underestimate of how widespread this pattern is in our set of species. For example, *Mus musculus* and *Apis mellifera*, which contain a high number of chromosomes, show no significant correlation between chromosome size and GC content (table 2). Yet, in both species, studies using recombination data inferred from genetic maps showed a significant positive correlation between local GC content and crossover rates (Beye et al. 2006; Khelifi et al. 2006; see table 1). In *M. musculus*, the absence of significant correlation between chromosome size and GC content can be explained by the lack of variance in chromosome size in that species (Meunier and Duret 2004). In *A. mellifera*, as in several other eukaryotes (e.g., *Schizosaccharomyces pombe*), chromosomes experience little or no crossover interference, and their mean recombination rate is therefore not correlated to their size, which explains that we do not observe any correlation between chromosome size and GC content in these species. Finally, it should be noted that the evolution of GC content is a slow process. If a genome has undergone recent chromosomal rearrangements, it might not show any significant correlation between chromosome size and GC content, simply because there was not enough time to establish the pattern (Duret and Arndt 2008). Given all these limitations of our test, it is remarkable that a majority of species (50–74% of all species with statistical power >50%) show correlations consistent with the predictions of the gBGC model.

Several species, however, do not fit into this general pattern: *Ciona instestinalis, C. neoformans*, *S. bicolor* and *T. brucei*. *Cryptococcus neoformans* is a species with evidence for gBGC from table 2 but not (or incompletely) from table 1. This can look surprising at first sight since we use

Table 1
Correlation between Recombination Rates and GC Content among Eukaryotes

Species	Eukaryotic groups[a]	Chromosome number	Genetic map[b]	Total GC/rec rates[c]	Chrom size/rec rates[c]
Saccharomyces cerevisiae	Unikonts	16	861	0.62* (*)	−0.6* (*)
Cryptococcus neoformans	Unikonts	13	285	0.04 ns (ns)	−0.14 ns (ns)
Monodelphis domestica	Unikonts	8	150	0.29 ns (ns)	−0.05 ns (ns)
Mus musculus	Unikonts	19	10195	0.68* (*)	−0.5* (ns)
Homo sapiens	Unikonts	22	28121	0.75*** (**)	−0.87*** (***)
Gallus gallus	Unikonts	27	9268	0.89*** (***)	−0.97*** (***)
Ciona intestinalis	Unikonts	13	276	−0.59* (ns)	0.21 ns (ns)
Caenorhabditis elegans	Unikonts	5	780	0.5 ns (ns)	−1* (*)
Drosophila melanogaster	Unikonts	4[d]	67	0.8 ns (ns)	−0.4 ns (ns)
Apis mellifera	Unikonts	16	2008	0.74* (*)	−0.35 ns (ns)
Trypanosoma brucei	Excavates	11	119	0.14 ns (ns)	−0.09 ns (ns)
Plasmodium falciparum	Chromal	14	3438	0.37 ns (ns)	−0.54* (ns)
Arabidopsis thaliana	Plantae	5	676	0 ns (ns)	−0.2 ns (ns)
Populus trichocarpa	Plantae	19	540	0.06 ns (ns)	−0.28 ns (ns)
Vitis vinifera	Plantae	19	515	−0.33 ns (ns)	−0.56* (*)
Oryza sativa	Plantae	12	1202	−0.18 ns (ns)	0.53 ns (ns)
Sorghum bicolor	Plantae	10	2029	0.5 ns (ns)	0.21 ns (ns)

[a]The eukaryotic groups relate to those shown in figure 1. Chromal, Chromalveolates.
[b]Number of markers in genetic maps.
[c]Values are Spearman correlation coefficients, then come *P* values: ns, nonsignificant, * <0.05, ** $<10^{-3}$, *** $<10^{-4}$ and *q* values (from FDR corrections for multiple tests) are indicated in parentheses.
[d]Here is indicated the number of chromosome arms instead of the number of chromosomes.

recombination data in table 1, which is a more direct way of testing for gBGC. However, this assumption is correct if recombination data are of high quality, which might not be the case for most of the species in table 1 with a small number of markers. Too few markers will tend to shorten genetic maps, underestimating recombination rates (other important parameters are the number of meioses analyzed, the distribution of markers along chromosomes). *Cryptococcus neoformans* and other species in table 1 may be in this situation. It is possible that in such species, chromosome length gives a better idea of the average chromosome-wide recombination rates, which could explain why we report comparatively more species showing evidence of gBGC in table 2 than in table 1. In *C. neoformans*, the use of two different strains for the available genetic map and the complete genome could be an additional problem for correlating GC content and recombination rates reliably. The conflicting results in *Ciona intestinalis* may also come from the poor-quality map found in this species (only 276 markers, see table 1). Using two genetic maps in *Plasmodium falciparum*, one from 1999 with 900 markers (Su et al. 1999) and a more recent one with 3,438 markers (Jiang et al. 2011), we found very different results (GC/recombination: −0.31 nonsignificant with the 1999 version map, 0.34 nonsignificant with the 2011 version map, Chromosome size/recombination: 0.23 nonsignificant with the 1999 version map, −0.54, $P < 0.05$ with the 2011 version map), which confirms that the quality of recombination data is critical. *Trypanosoma brucei* shows a significant positive correlation between chromosome size and GC content (fig. 2*b*). However, it turns out that, for an unknown reason, chromosome size is not a good proxy for recombination rate in this species: the two parameters are not correlated ($\rho = -0.09$; $P = 0.797$, see table 1). Table 1 reveals that GC content correlates positively with recombination rates in *T. brucei* ($\rho = 0.14$), although not significantly ($P = 0.694$). It thus appears that *T. brucei* is not an exception to the general pattern consistent with gBGC. Again, a better map in this species would help understand more clearly the relationships between GC content, chromosome size and recombination rates (there are only 119 markers in this species, see table 1). In *S. bicolor*, GC3 correlates strongly with chromosome size in a positive manner (table 2). We do not have explanations for this significant correlation, which is not in agreement with gBGC. *Sorghum bicolor* seems therefore to represent a true exception to the general pattern.

In conclusion, we found 17 species with a significant correlation between chromosome-wide GC content and chromosome size, as a rough proxy for recombination rate. Most of them (15/17) showed a negative correlation, consistent with the gBGC model. Our results were unaltered when considering GC3, which rules out the insertion of transposable elements as a general explanation for the observed pattern. Other explanations are of course possible (mutational biases, selection on GC content). In species where these various hypotheses have been tested, gBGC has always come out as the most likely explanation (reviewed in Marais 2003; Duret and Galtier 2009a). More work will be needed, however, to test these alternative explanations and firmly establish

Table 2
Correlation between Chromosome Size and GC Content among Eukaryotes

Species	Eukaryotic groups[a]	Chromosome number	Mean GC content (%)	Statistical power[b] (%)	GC total/chrom size[c]	GC3/chrom size[c]
Encephalitozoon cuniculi	Unikonts	11	47	41	0.3 ns (ns)	0.06 ns (ns)
Schizosaccharomyces pombe	Unikonts	3	36	0	−0.5 ns (ns)	−0.5 ns (ns)
Saccharomyces cerevisiae	Unikonts	16	38	70	−0.83*** (**)	−0.87*** (***)
Candida glabrata	Unikonts	13	39	52	−0.69* (*)	−0.71* (*)
Pichia stipitis	Unikonts	8	41	25	0.24 ns (ns)	0.71 ns (ns)
Yarrowia lipolytica	Unikonts	6	49	14	0.77 ns (ns)	−0.09 ns (ns)
Aspergillus fumigatus	Unikonts	8	50	25	0.71 ns (ns)	0.26 ns (ns)
Magnaporthe grisea	Unikonts	7	52	22	−0.11 ns (ns)	0.07 ns (ns)
Gibberella zeae	Unikonts	4	48	0	0.4 ns (ns)	0.4 ns (ns)
Ustilago maydis	Unikonts	23	54	100	−0.46* (ns)	−0.47* (*)
Cryptococcus neoformans	Unikonts	14	49	58	−0.33 ns (ns)	−0.72* (*)
Coprinopsis cinerea	Unikonts	13	52	52	−0.91*** (***)	−0.68* (*)
Monodelphis domestica	Unikonts	9	38	28	−0.1 ns (ns)	−0.07 ns (ns)
Mus musculus	Unikonts	20	42	99	−0.28 ns (ns)	−0.26 ns (ns)
Homo sapiens	Unikonts	23	41	100	−0.57* (*)	−0.54* (*)
Gallus gallus	Unikonts	29	41	100	−0.93*** (***)	−0.97*** (***)
Ciona intestinalis	Unikonts	13	36	52	0.2 ns (ns)	0.29 ns (ns)
Caenorhabditis elegans	Unikonts	6	35	14	−0.54 ns (ns)	−0.26 ns (ns)
Drosophila melanogaster	Unikonts	5[d]	42	7	−0.3 ns (ns)	−0.5 ns (ns)
Apis mellifera	Unikonts	16	35	70	−0.03 ns (ns)	−0.16 ns (ns)
Dictyostelium discoideum	Unikonts	6	22	14	−0.94* (*)	−0.6 ns (ns)
Trypanosoma brucei	Excavates	11	46	41	0.73* (*)	0.48 ns (ns)
Leishmania major	Excavates	36	60	100	−0.85*** (***)	−0.82*** (***)
Guillardia theta	Chromal	3	26	0	−1 ns (ns)	−1 ns (ns)
Paramecium tetraurelia	Chromal	114	28	100	−0.84*** (***)	−0.89*** (***)
Plasmodium falciparum	Chromal	14	19	58	−0.8** (*)	−0.77* (*)
Cryptosporidium parvum	Chromal	8	30	25	0.1 ns (ns)	−0.1 ns (ns)
Thalassiosira pseudonana	Chromal	23	47	100	−0.06 ns (ns)	−0.87*** (***)
Phaeodactylum tricornutum	Chromal	33	49	100	−0.18 ns (ns)	−0.46* (*)
Ostreococcus lucimarinus	Plantae	19	60	94	−0.72** (*)	−0.66* (*)
Micromonas pusilla	Plantae	15	64	63	−0.83** (**)	−0.42 ns (ns)
Arabidopsis thaliana	Plantae	5	36	7	−0.2 ns (ns)	−0.3 ns (ns)
Vitis vinifera	Plantae	19	34	94	0.37 ns (ns)	−0.31 ns (ns)
Populus trichocarpa	Plantae	19	33	94	0.22 ns (ns)	0.36 ns (ns)
Oryza sativa	Plantae	12	44	46	0.47 ns (ns)	0.48 ns (ns)
Sorghum bicolor	Plantae	10	44	36	0.3 ns (ns)	0.68* (*)

[a]The eukaryotic groups relate to those shown in figure 1. Chromal, Chromalveolates.
[b]Statistical power for chromosome number ≥23 is set to 100%.
[c]Values are Spearman correlation coefficients, then come *P* values: ns, nonsignificant, * <0.05,** $<10^{-3}$,*** $<10^{-4}$ and *q* values (from FDR corrections for multiple tests) are indicated in parentheses.
[d]Here is indicated the number of chromosome arms instead of the number of chromosomes.

gBGC in the species where we report data consistent with gBGC for the first time. Figure 1 shows, in our set of 36 eukaryotes, the species with a positive correlation between GC content and recombination rates (measured directly or using chromosome size as a proxy), consistent with gBGC. Remarkably, this correlation is found in all four major eukaryotic groups studied, which suggests gBGC is widespread in eukaryotes. This is in agreement with a recent study using GC content of ribosomal DNA as a proxy for gBGC, in which gBGC was inferred in several distantly related eukaryotes (Escobar et al. 2011). Firm evidence for gBGC is only available for a handful of species (yeasts and mammals) and our work suggests that gBGC should be further studied in many more species, where it could have important effects on genome evolution.

Materials and Methods

Genome Data

We selected species for which a complete genome assembly, anchored on chromosomes, was available. Animal species are clearly over-represented in public databases. As gBGC is

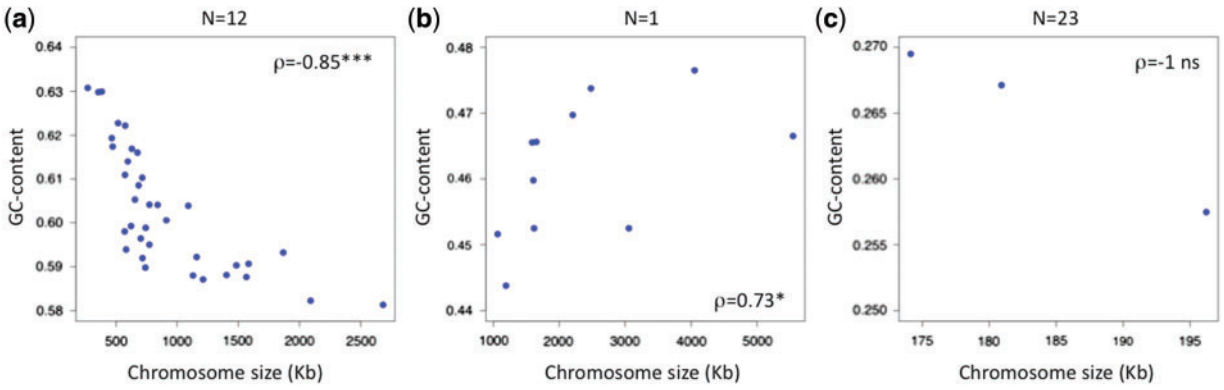


Fig. 2.—Examples of relationships between chromosome size and total GC content. (*a*) *L. major*. (*b*) *T. brucei*. (*c*) *G. theta*. The values above the plots indicate the number of similar observations that were made among the 36 species (e.g., $N = 12$ for [a] means 12 significant positive correlations). ρ = Spearman coefficient. Statistical significance: ns, nonsignificant, * <0.05, ** $<10^{-3}$, *** $<10^{-4}$.

already established in animals, we only selected a subset of species representing the main animal groups. Genome data were extracted from Hogenom version 3 (17 species [Penel et al. 2009]), the NCBI website (15 species, http://www.ncbi.nlm.nih.gov/), and the JGI website (4 species, http://www.jgi.doe.gov/). For the *Paramecium* genome, we selected the scaffolds that were at least chromosomal arms (Gout J-F, personal communications). The relationship between chromosome size and recombination rate only stands for recombining chromosomes and we therefore removed all the nonrecombining chromosomes (chromosomes 4 from *Drosophila melanogaster*, 2 and 18 from *Ostreococcus lucimarinus*, 1 and 17 from *M. pusilla*, Y and W chromosomes from mammals and chicken, respectively). For our 36 species, we thus had chromosome sizes and sequences to estimate the GC content.

Recombination Data

We got recombination data for *C. elegans* directly from MareyMap (Rezvoy et al. 2007), *D. melanogaster* from Flybase (http://flybase.org) and *Saccharomyces cerevisiae* from http://www.yeastgenome.org/pgMaps/pgI.shtml. Recombination data for other species was obtained from specific papers: *M. musculus* (Cox et al. 2009), *Homo sapiens* (Matise et al. 2007), *Gallus gallus* (Groenen et al. 2009), *Monodelphis domestica* (Samollow et al. 2007), *A. mellifera* (Beye et al. 2006), *T. brucei* (Cooper et al. 2008), *P. falciparum* (Jiang et al. 2011), *Arabidopsis thaliana* (Singer et al. 2006), *C. intestinalis* (Kano et al. 2006), *S. bicolor* (Mace et al. 2009), *Populus trichocarpa* (Yin et al. 2004), *Vitis vinifera* (Doligez et al. 2006), *Oryza sativa* (Muyle et al. 2011), and *C. neoformans* (Marra et al. 2004). The number of chromosomes indicated in table 1 may differ from the true chromosome number: the X and Z chromosomes were excluded from this analysis because they recombine only in one sex, and recombination patterns are thus different from those in the autosomes, and the recombination data are not available for some chromosomes (for instance, chromosome 10 for *C. neoformans*). The recombination rates were computed by dividing the genetic map length of each chromosome by its physical size (in bp), and are thus chromosomal-averaged estimates.

GC Content Analysis

The total GC content was computed using whole-chromosome sequences. The GC content at third codon position (GC3) was computed by collecting all the available CDS from a genome (extracting CDS from Hogenom or Ensembl, or using CDS files from JGI or Broad Institute). For both total GC content and GC3 estimates, ambiguous nucleotides were excluded. Chromosome-averaged GC values were then computed. *R* was used to obtain bilateral Spearman coefficients of correlation, *P* values, and *q* values (*P* values corrected for multiple testing using the false discovery rate method). The combined analysis was performed by first getting the *P* values (*P*) from unilateral tests on Spearman coefficients in order to test for a general trend for a negative correlation between GC content and chromosome size (null hypothesis: GC content and chromosome size are not correlated negatively). The sum of the $-2 * \log(P)$ for all species follows a chi-squared distribution with $2n$ degrees of freedom, n being the number of species, which gave the *P* value of the combined analysis (Sokal and Rohlf 2012).

Statistical Power Analysis

To estimate the power of our approach according to the number of chromosomes (N) in a given genome, we performed the following test: we took the human genome (for which there is clear evidence of gBGC and which shows a significant negative correlation between chromosome size and GC content) and we asked what would be the probability

to detect a significant correlation if this genome only contained *N* chromosomes. We thus randomly sampled *N* human chromosomes, computed the Spearman coefficient between their size and GC content, repeated this for all the possible combinations (up to 50,000 samples) and measured the fraction of significant Spearman correlations in the simulated data using R. We took this fraction as the statistical power of our test for a given number of chromosomes *N*. The same was done using *S. cerevisiae* as a reference.

Acknowledgments

We thank Jean-François Gout for his help with the *Paramecium* genome data. This work was supported by the Centre National de la Recherche Scientifique and the Agence Nationale de la Recherche (ANR-08-GENM-036-01).

Literature Cited

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol. 7:e26.
Beye M, et al. 2006. Exceptionally high levels of recombination across the honey bee genome. Genome Res. 16:1339–1344.
Birdsell JA. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. Mol Biol Evol. 19:1181–1197.
Bradnam KR, Seoighe C, Sharp PM, Wolfe KH. 1999. G+C content variation along and among Saccharomyces cerevisiae chromosomes. Mol Biol Evol. 16:666–675.
Capra JA, Pollard KS. 2011. Substitution patterns are GC-biased in divergent sequences across the metazoans. Genome Biol Evol. 3:516–527.
Cooper A, et al. 2008. Genetic analysis of the human infective trypanosome Trypanosoma brucei gambiense: chromosomal segregation, crossing over, and the construction of a genetic map. Genome Biol. 9:R103.
Copenhaver GP, Browne WE, Preuss D. 1998. Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads. Proc Natl Acad Sci U S A. 95:247–252.
Cox A, et al. 2009. A new standard genetic map for the laboratory mouse. Genetics 182:1335–1344.
Doligez A, et al. 2006. An integrated SSR map of grapevine based on five mapping populations. Theor Appl Genet. 113:369–382.
Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. 4:e1000071.
Duret L, et al. 2008. Analysis of sequence variability in the macronuclear DNA of Paramecium tetraurelia: a somatic view of the germline. Genome Res. 18:585–596.
Duret L, Galtier N. 2009a. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet. 10:285–311.
Duret L, Galtier N. 2009b. Comment on "Human-specific gain of function in a developmental enhancer." Science 323:714; author reply 714.
Escobar JS, et al. 2010. An integrative test of the dead-end hypothesis of selfing evolution in Triticeae (Poaceae). Evolution 64:2855–2872.
Escobar JS, Glemin S, Galtier N. 2011. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. Mol Biol Evol. 28:2561–2575.
Eyre-Walker A. 1993. Recombination and mammalian genome evolution. Proc Biol Sci. 252:237–243.
Galtier N. 2003. Gene conversion drives GC content evolution in mammalian histones. Trends Genet. 19:65–68.
Galtier N, Bazin E, Bierne N. 2006. GC-biased segregation of noncoding polymorphisms in Drosophila. Genetics 172:221–228.
Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. 23:273–277.
Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. 25:1–5.
Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911.
Groenen MA, et al. 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. Genome Res. 19:510–519.
Haudry A, et al. 2008. Mating system and recombination affect molecular evolution in four Triticeae species. Genet Res. 90:97–109.
International Chicken Genome Sequencing Consortium. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695–716.
Jancek S, Gourbiere S, Moreau H, Piganeau G. 2008. Clues about the genetic basis of adaptation emerge from comparing the proteomes of two Ostreococcus ecotypes (Chlorophyta, Prasinophyceae). Mol Biol Evol. 25:2293–2300.
Jiang H, et al. 2011. High recombination rates and hotspots in a Plasmodium falciparum genetic cross. Genome Biol. 12:R33.
Kaback DB. 1996. Chromosome-size dependent control of meiotic recombination in humans. Nat Genet. 13:20–21.
Kaback DB, Barber D, Mahon J, Lamb J, You J. 1999. Chromosome size-dependent control of meiotic reciprocal recombination in Saccharomyces cerevisiae: the role of crossover interference. Genetics 152:1475–1486.
Kano S, Satoh N, Sordino P. 2006. Primary genetic linkage maps of the ascidian, Ciona intestinalis. Zoolog Sci. 23:31–39.
Keeling PJ, et al. 2005. The tree of eukaryotes. Trends Ecol Evol. 20: 670–676.
Khelifi A, Meunier J, Duret L, Mouchiroud D. 2006. GC content evolution of the human and mouse genomes: insights from the study of processed pseudogenes in regions of different recombination rates. J Mol Evol. 62:745–752.
Mace ES, et al. 2009. A consensus genetic map of sorghum that integrates multiple component maps and high-throughput Diversity Array Technology (DArT) markers. BMC Plant Biol. 9:13.
Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. Nature 454:479–485.
Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. Trends Genet. 19:330–338.
Marais G, Charlesworth B, Wright SI. 2004. Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*. Genome Biol. 5:R45.
Marais G, Mouchiroud D, Duret L. 2001. Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes. Proc Natl Acad Sci U S A. 98:5688–5692.
Marais G, Mouchiroud D, Duret L. 2003. Neutral effect of recombination on base composition in Drosophila. Genet Res. 81:79–87.
Marra RE, et al. 2004. A genetic linkage map of Cryptococcus neoformans variety neoformans serotype D (Filobasidiella neoformans). Genetics 167:619–631.
Matise TC, et al. 2007. A second-generation combined linkage physical map of the human genome. Genome Res. 17:1783–1786.
Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. Mol Biol Evol. 21:984–990.
Mikkelsen TS, et al. 2007. Genome of the marsupial Monodelphis domestica reveals innovation in non-coding sequences. Nature 447: 167–177.

Montoya-Burgos JI, Boursot P, Galtier N. 2003. Recombination explains isochores in mammalian genomes. Trends Genet. 19: 128–130.
Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glemin S. 2011. GC-biased gene conversion and selection affect GC content in the Oryza genus (rice). Mol Biol Evol. 28:2695–2706.
Nabholz B, Kunstner A, Wang R, Jarvis ED, Ellegren H. 2011. Dynamic evolution of base composition: causes and consequences in avian phylogenomics. Mol Biol Evol. 28:2197–2210.
Penel S, et al. 2009. Databases of homologous gene families for comparative genomics. BMC Bioinformatics 10(Suppl 6), S3.
Ratnakumar A, et al. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. Philos Trans R Soc Lond B Biol Sci. 365:2571–2580.
Rezvoy C, Charif D, Gueguen L, Marais GA. 2007. MareyMap: an R-based tool with graphical interface for estimating recombination rates. Bioinformatics 23:2188–2189.
Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. Genome Res. 20: 1001–1009.
Samollow PB, et al. 2007. A microsatellite-based, physically anchored linkage map for the gray, short-tailed opossum (Monodelphis domestica). Chromosome Res. 15:269–281.
Singer T, et al. 2006. A high-resolution map of Arabidopsis recombinant inbred lines by whole-genome exon array hybridization. PLoS Genet. 2:e144.
Sokal R, Rohlf F. 2012. Biometry: the principles and practices of statistics in biological research. New York: W.H. Freeman & Co Ltd.
Spencer CC, et al. 2006. The influence of recombination on human genetic diversity. PLoS Genet. 2:e148.
Su X, et al. 1999. A genetic map and recombination parameters of the human malaria parasite Plasmodium falciparum. Science 286: 1351–1353.
Webster MT, Hurst LD. 2011. Direct and indirect consequences of meiotic recombination: implications for genome evolution. Trends Genet. 28:101–109.
Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA. 2004. Large-scale heterospecific segregation distortion in Populus revealed by a dense genetic map. Theor Appl Genet. 109:451–463.

Associate editor: Laurence Hurst