Genomic footprints of speciation with gene flow

Kiwoong Nam[1]

Speciation can be hampered by gene flow between a pair of diverging populations. A large number of theoretical speciation models have proposed certain conditions by which speciation with gene flow may occur, while we still have limited empirical support. Since whole genome sequences have ample footprints of evolutionary processes of speciation, genomic analyses can be useful approaches to infer the past history of speciation with gene flow. Here, we review which patterns are expected in whole genome sequences in the presence of speciation with gene flow.

Keyword: genome hitchhikng, Speciation with gene flow, sympatric speciation

[1]*DGIMI, Univ Montpellier, INRAE, Montpellier, France*

*__Corresponding author:__ ki-woong.nam@inrae.fr

INTRODUCTION

Speciation occurs in the presence of reproductive barriers between a pair of diverging populations. If populations are geographically separated, then divergence may occur readily by natural selection or genetic drift because gene flow can be effectively suppressed between populations. However, if such a geographic separation is absent, a substantial proportion of individuals can migrate between populations. Then, these migrating individuals cause genetic exchanges between populations. This exchange, gene flow, hampers genetic differentiation, making speciation difficult. How speciation with gene flow occurs has been one of the most debating issues in speciation biology (Bolnick and Fitzpatrick 2007). This topic has been expressed as 'sympatric speciation' (Barluenga et al. 2006). But, it was often unclear whether a whole process of speciation occurs in sympatry. For example, Barluenga et al. (2006) suggested that cichlid fishes in Nicaraguan crater lake experienced sympatric speciation. However, it has been unclear whether these species had a certain period of geographic separation (Schliewen et al. 1994). A growing number of speciation biologists believe that, rather than testing absolute sympatry during the whole process of speciation, which could be extreme cases of biological reality, we will get better insights into speciation by focusing on how speciation occurs despite the existence of gene flow between sympatric populations (please see Mallet 2005; Fitzpatrick et al. 2008). Therefore, the focus was moved from sympatry to gene flow, to understand how speciation occurred without absolute geographic separation (Fitzpatrick et al. 2008).

Genome sequences have accumulated footprints of various types of evolutionary processes, including speciation. Therefore, whole genome analyses have been suggested to be powerful tools to unveil evolutionary processes of speciation with gene flow, especially in cases with incipient speciation. Thanks to the advances in sequencing technology, large-scale whole genome sequencing can be performed relatively easily these days. In this review, we will discuss footprints of speciation with gene flow in the genome.

MAIN

Difficulties in speciation with gene flow

Felsenstein (1981) showed in his paper that recombination hampers speciation in the presence of gene flow. He considers a situation where a pair of populations are under two diverging evolutionary forces, assortative mating and ecological divergent selection, which are controlled by two different loci. Recombination generates all possible allelic combinations of these two loci. Then, the direction of divergence is determined by the relative strength between assortative matings and ecological selection.

For example, by the ecological disruptive selection, AA and aa will be selected over Aa at locus 1. Assortative mating will increase the frequency of BB and bb over Bb at locus 2. Recombination between locus 1 and 2 will generate the following nine allelic combinations; AA BB, Aa BB, aa BB, AA Bb, Aa Bb, aa Bb, AA bb, Aa bb, and aa bb. By the ecological disruptive selection, AA BB, aa BB, AA Bb, aa Bb, AA bb, and aa bb will be selected. However, among these six combinations, the proportions of AA Bb and aa Bb will be decreased by assortative matings. Instead, assortative mating will increase the proportion of individuals with Aa BB and Aa bb even though these two combinations will cause reduced ecological fitness. Therefore, divergence by ecological divergent disruptive selection will be interfered with by assortative matings. Thus, speciation will be completed only when this interference is overcome. So far, more than 100 models of speciation with gene flow have been proposed depending upon the condition that the interference by recombination is effectively reduced (Gavrilets 2014).

If a single trait determines both assortative matings and ecological divergent selection, for example, then recombination does not have interfering effects because there is only one trait concerning ecological divergent selection and assortative mating. For example, in the *Rhagoletis pomonella* sibling-species complex (common name: apple maggot flies), the choice of host plants is under ecological divergent selection because the host plant provides nutrients (Linn et al. 2004) as well as mating places (Feder et al. 1994). This type of trait was coined ‘magic trait’ (Servedio et al. 2011).

Even though we have now ample empirical cases of speciation in the presence of gene flow, we have still limited knowledge of the evolutionary conditions that the homogenizing effect of gene flow overcome in these cases. Interestingly, speciation with gene flows was reported mostly from phytophagous insects (Berlocher and Feder 2002). Thus, phytophagous insects could be optimal organisms to unveil the process of speciation with gene flow.

The differentiation in the whole genome sequences

A process of speciation involves the genetic differentiation across a whole genome between a pair of speciating taxa. In the absence of geographic separation, while loci under divergent selection are genetically differentiated, effectively neutrally evolving genetic elements are constantly homogenized by gene flow. Therefore, speciation with gene flow may be complete only when divergent selection affects across the whole genome sequences.

For a long time, this whole genome differentiation has been believed to be initiated from a small number of loci under divergent selection, and genetically differentiated loci are progressively enlarged by following divergent selection until whole genome sequences are differentiated. This speciation model was coined 'Genic view of speciation (Wu 2001). This model assumes the uniform direction of divergent selection throughout the whole speciation process, which may take millions of years. However, it might not be realistic in many species that environmental changes do not cause an altered direction of divergence by selection during the whole process of speciation.

Theoretical studies have demonstrated, however, that whole genome sequences can be differentiated at the beginning of a speciation process in the presence of extremely strong divergent selection or atypical genome structures. It should be noted that, here, whole genome differentiation means a status that genetic differentiation is observed across the whole genome even though the level of differentiation could be low (e.g., low F_{ST}), instead of the status that whole genome sequences have complete genetic differentiation (e.g., F_{ST} = 1) (Nam et al. 2020). If selection is sufficiently strong (e.g., $s > m$ (Flaxman et al. 2014) or $s > r$ (Barton 1979), where s, m, and r are selection coefficient, migration rate, and recombination rate, respectively), the diverging effect of selection dominates the homogenizing effect of gene flow, thus whole genome differentiation may occur. In addition, atypical genome structures have been proposed to be a driver of whole genome differentiation, such as a tight physical linkage between selectively targeted loci and chromosomal rearrangement (Feder, Nosil, and Flaxman 2014) or clusters of reproductive isolation loci within a genome (Via and West 2008; Via 2012).

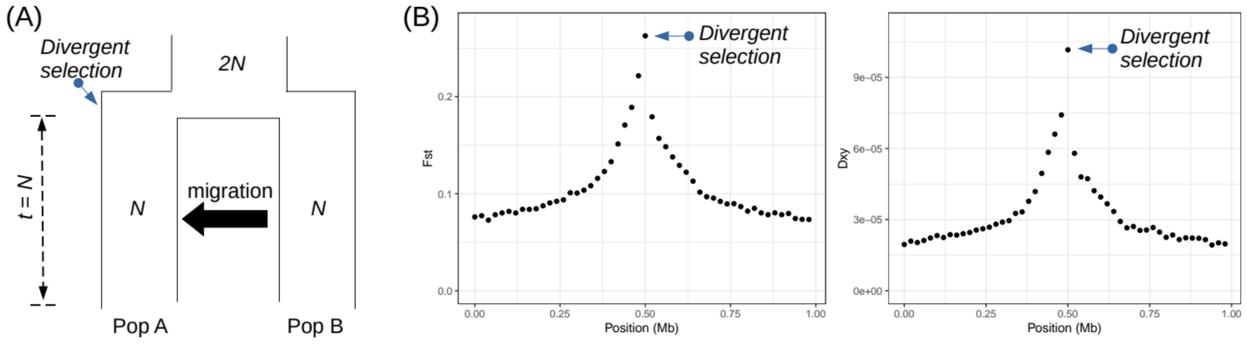

Figure 1. The expected pattern of F_{ST} and D_{XY} in the presence of divergent selection. (A) The evolutionary scenario used for forward simulation using slim4 (Haller and Messer 2019). An ancestral population with a population size equal to $2N$ split into two populations, Pop A and Pop B, N generations ago. The population sizes of Pop A and Pop B were N. Unidirectional migration occurs from Pop B to Pop A with the rate of gene flow equal to 0.001. The recombination rate, mutation rate, and population size were 1.19×10^{-8}, 1.2×10^{-8}, and 3,100, respectively, which were chosen from the human condition (Kong et al. 2002; Tenesa et al. 2007; Campbell et al. 2012). The simulated DNA sequence was 1Mb in length, and Pop A experienced divergent selection with s = 0.02 in the middle of the sequence. In total, 500 independent

simulations were performed and F_{ST} and D_{XY} were averaged out. (B) Selectively targeted locus has increased F_{ST} and D_{XY}.

Other theoretical studies showed that mild divergent selection alone can be sufficient for whole genome differentiation. If the number of selectively targeted loci is sufficiently high, the migration rate can be effectively reduced across the whole genome because the combined effect of selection can be sufficiently high to suppress the migration rate across the whole genome (termed genome hitchhiking model (Feder and Nosil 2010; Feder et al. 2012)), depending upon the relative strength of selection to the rate of gene flow. Alternatively, if a very large number of loci are targeted by selective sweeps, the average distance from a locus to the nearest selectively targeted locus decreases, and whole genome differentiation may occur by physical linkage to the targets (Barton and Bengtsson 1986). Therefore, mild positive selection can be theoretically sufficient for whole genome differentiation. In other words, the number of selectively targeted loci can be a single factor determining the potential for whole genome differentiation.

If divergent positive selection involves polygenic adaptation with a quantitative trait, then adaptive evolution may occur by subtle changes in allele frequencies of a large number of loci (Barghi et al. 2020). Since most theoretical speciation models concern a situation of selective sweeps, which correspond that divergent selection occurs by the fixation of beneficial mutations and that genetically linked targeted loci are consequently genetically differentiated, it is underexplored how polygenic adaptation may affect speciation with gene flow. Future studies might be able to address whether polygenic adaptation may cause the reduction of effective migration rate across the whole genome. In this case, the link between polygenic adaptation and the genome hitchhiking could be stated as well.

The fall armyworm (*Spodoptera frugiperda*, Lepidoptera) is composed of two sympatric strains with differentiated host plants. These two strains have a very low level of genetic differentiation (F_{ST} = 0.0174), but more than 99% of 200kb windows in the 400Mb genome have F_{ST} higher than 0, implying that whole genome sequences have been differentiated at the beginning of potential incipient speciation (Nam et al. 2020). In apple maggot files, selectively targeted loci are reported to be widespread across a genome as well (Michel et al. 2010). However, it is still unclear whether mild divergent selection alone has been a driver of whole genome differentiation. As more empirical cases of whole genome differentiation are reported, the role of divergent selection could be clearly identified.

Whole genomic differentiation at the beginning of speciation may provide a condition for accelerated speciation by following events of divergent selection. Theoretical studies demonstrated that if the number of targets is higher than a certain threshold, then targeted loci have a synergistic effect in increasing linkage disequilibrium among targets, thus genomic differentiation is consequently increased (Barton 2010; Flaxman et al. 2014). The non-linear dynamics of genomic differentiation according to the number of occurred selection events were termed genome-wide congealing (Feder, Nosil, Wacholder, et al. 2014). It should be noted that any diversifying factors, including divergent selection, background selection, and assortative mating (Kopp et al. 2017), may also contribute to genome-wide congealing. Empirical reports of this non-linear pattern are still limited. Future studies include model species from which this non-linear relationship can be tested in the context of the speciation continuum.

Genomic analyses to identify divergent selection

Genome-scan to identify loci targeted by selective sweeps is a useful tool to study the role of divergent selection in speciation with gene flow. Targets of selective sweeps have distinct patterns, such as increased genetic differentiation (Chen et al. 2010) (i.e., high F_{ST}), an increased proportion of rare alleles (Kim 2006; Pavlidis et al. 2013) (negative Tajima's D), decreased genetic diversity (decreased π), increased linkage disequilibrium (high r^2) (Kim and Nielsen 2004), and increased length of shared haplotypes (high iHS) (Voight et al. 2006). Thus, genes under selective sweeps can be identified based on these statistics.

There are two main challenges to this approach. First, background selection generates the same patterns as selective sweeps (Comeron 2014). The single difference between selective sweeps and background selection is no more than the extent of the patterns. For example, selective sweeps cause a reduction in genetic diversity much more severely than background selection. In this case, targets of selective sweeps could be identified from loci where the reduction of genetic diversity cannot be explained only by background selection (Nam et al. 2015). However, in most cases, expected F_{ST}, Tajima's D, π, r^2, or iHS in the presence of background could be reliably inferred only when we know local variations in the gene density, the recombination rate, and the mutation rates across a genome, as well as past demographic history. Alternatively, statistical outliers can be regarded as a target of selective sweeps with human-determined criteria based on the assumption that most genomic loci are not targeted by selective sweeps. For example, loci with the 99% highest F_{ST} could be considered as being targeted by selective sweeps. However, this human criterion is often hard to justify. Recently, machine learning is often used to identify selective sweeps (Kern and Schrider 2018; Hejase et al. 2022).

The second challenge is the low resolution of selection scans. Selective sweeps generate their footprints not only on the target genes but also on other genes that are genetically linked to the target. Therefore, it is often hard to pinpoint the target precisely among several genes within a locus under selective sweeps. It is important to note that only one genetic variation within the same linkage disequilibrium will be selected because of Hill-Roberson interference (McVean and Charlesworth 2000; Castellano et al. 2016) unless multiple genetic variations in a single linkage disequilibrium have additive or multiplicate effects during selection. The resolution of selection scans can be dramatically increased by phasing, a bioinformatics process of identifying genetic variations with the same parental origins, even up to a single base pair (Grossman et al. 2013). Third-generation sequencing (PacBio or Oxford nanopore) can be particularly useful for phasing with long reads. Technological advancement of third-generation sequencing may enable to obtain fully phased genetic variation across a whole genome from population data with a realistic cost in a small lab.

The increased level of absolute differentiation should also be used to identify divergent selection causing reproductive isolation, in addition to the classical approaches to identify selective sweeps (Cruickshank and Hahn 2014). F_{ST} is a proportion of variance explained by the difference between populations by definition (Weir and Cockerham 1984). If a locus has F_{ST} higher than the genomic average, it only means that this locus has an increased level of non-random distribution of genetic variation according to populations. This non-randomness can be increased only by background selection. Therefore, increased F_{ST} in a locus itself does not necessarily mean the presence of divergent selection.

D_{XY} is the probability that one individual from one population has different a sequence from one individual from another population. If divergent selection caused reproductive isolation by a reduction in gene flow, the corresponding targeted locus is expected to have increased D_{XY} because shared genetic variations between populations will be removed by divergent selection (Hejase et al. 2020). In short, selectively targeted loci causing reproductive isolation should have both increased D_{XY} and F_{ST} (Fig. 1)

Bioinformatics genome scan is often followed by a functional genomics test to verify the function of the identified genes or to test their adaptive role. Knock-out through CRISPR/CAS9 (Mathyer et al. 2021) is increasingly used for this purpose. According to the technical advancement of functional genomics, functional genomics tests are expected to be more realistic in a growing number of non-model species. This functional test is gradually becoming almost a mandatory step in a typical evolutionary genomics study. Thus, evolutionary geneticists have the necessity for regular check-ups of these technics for studies based on population genomics analyses.

CONCLUSIONS

The advancements in theoretical speciation biology indeed demonstrated that speciation with gene flow may occur through numerous models of speciation. In addition, recently developed technique of sequencing and functional genomics enable reliable identification of genes under divergent selection. Roughly speaking, studies of evolutionary processes of speciation with gene flow are limited by our ability to identify optimal species, rather than available skills or techniques. If we find suitable species, it might be relatively straightforward to test the role of whole genome differentiation, the effects of mild or strong divergent selection, and the functional contribution of divergent selection during the speciation process with gene flow.

ACKNOWLEDGMENT

This work (ID 1702-018) was publicly funded through ANR (the French National Research Agency) under the "Investissements d'avenir" programme with the reference ANR-10-LABX-001-01 Labex Agro and coordinated by Agropolis Fondation under the frame of I-SITE MUSE (ANR-16-IDEX-0006).

REFERENCES

Barghi N, Hermisson J, Schlötterer C. 2020. Polygenic adaptation: a unifying framework to understand positive selection. Nat. Rev. Genet. 21:769-781.

Barluenga M, Stölting KN, Salzburger W, Muschick M, Meyer A. 2006. Sympatric speciation in Nicaraguan crater lake cichlid fish. Nature 439:719-723.

Barton N, Bengtsson BO. 1986. The barrier to genetic exchange between hybridising populations. Heredity 57:357-376.

Barton NH. 1979. Gene flow past a cline. Heredity 43:333-339.

Barton NH. 2010. What role does natural selection play in speciation? Philos. Trans. R. Soc. B Biol. Sci. 365:1825-1840.

Berlocher SH, Feder JL. 2002. Sympatric speciation in phytophagous insects: moving beyond controversy? Annu. Rev. Entomol. 47:773-815.

Bolnick DI, Fitzpatrick BM. 2007. Sympatric speciation: models and empirical evidence. Annu. Rev. Ecol. Evol. Syst. 38:459-487.

Campbell CD, Chong JX, Malig M, Ko A, Dumont BL, Han L, Vives L, O'Roak BJ, Sudmant PH, Shendure J, et al. 2012. Estimating the human mutation rate using autozygosity in a founder population. Nat. Genet. 44:1277-1281.

Castellano D, Coronado-Zamora M, Campos JL, Barbadilla A, Eyre-Walker A. 2016. Adaptive Evolution Is Substantially Impeded by Hill-Robertson Interference in Drosophila. Mol. Biol. Evol. 33:442-455.

Chen H, Patterson N, Reich D. 2010. Population differentiation as a test for selective sweeps. Genome Res. 20:393-402.

Comeron JM. 2014. Background Selection as Baseline for Nucleotide Variation across the Drosophila Genome. PLOS Genet. 10:e1004434.

Cruickshank TE, Hahn MW. 2014. Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. Mol. Ecol. 23:3133-3157.

Feder JL, Gejji R, Yeaman S, Nosil P. 2012. Establishment of new mutations under divergence and genome hitchhiking. Philos. Trans. R. Soc. B Biol. Sci. 367:461-474.

Feder JL, Nosil P. 2010. The efficacy of divergence hitchhiking in generating genomic islands during ecological speciation. Evol. Int. J. Org. Evol. 64:1729-1747.

Feder JL, Nosil P, Flaxman SM. 2014. Assessing when chromosomal rearrangements affect the dynamics of speciation: implications from computer simulations. Front. Genet. 5:295.

Feder JL, Nosil P, Wacholder AC, Egan SP, Berlocher SH, Flaxman SM. 2014. Genome-wide congealing and rapid transitions across the speciation continuum during speciation with gene flow. J. Hered. 105:810-820.

Feder JL, Opp SB, Wlazlo B, Reynolds K, Go W, Spisak S. 1994. Host fidelity is an effective premating barrier between sympatric races of the apple maggot fly. Proc. Natl. Acad. Sci. U. S. A. 91:7990-7994.

Felsenstein J. 1981. Skepticism towards Santa Rosalia, or why are there so few kinds of animals?

Evolution 35:124-138.

Fitzpatrick BM, Fordyce JA, Gavrilets S. 2008. What, if anything, is sympatric speciation? J. Evol. Biol. 21:1452-1459.

Flaxman SM, Wacholder AC, Feder JL, Nosil P. 2014. Theoretical models of the influence of genomic architecture on the dynamics of speciation. Mol. Ecol. 23:4074-4088.

Gavrilets S. 2014. Models of Speciation: Where Are We Now? J. Hered. 105:743-755.

Grossman SR, Andersen KG, Shlyakhter I, Tabrizi S, Winnicki S, Yen A, Park DJ, Griesemer D, Karlsson EK, Wong SH, et al. 2013. Identifying Recent Adaptations in Large-scale Genomic Data. Cell 152:703-713.

Haller BC, Messer PW. 2019. Evolutionary Modeling in SLiM 3 for Beginners. Mol. Biol. Evol. 36:1101-1109.

Hejase HA, Mo Z, Campagna L, Siepel A. 2022. A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph. Mol. Biol. Evol. 39:msab332.

Hejase HA, Salman-Minkov A, Campagna L, Hubisz MJ, Lovette IJ, Gronau I, Siepel A. 2020. Genomic islands of differentiation in a rapid avian radiation have been driven by recent selective sweeps. Proc. Natl. Acad. Sci. U. S. A. 117:30554-30565.

Kern AD, Schrider DR. 2018. diploS/HIC: An Updated Approach to Classifying Selective Sweeps. G3 Genes Genomes Genet. 8:1959-1970.

Kim Y. 2006. Allele frequency distribution under recurrent selective sweeps. Genetics 172:1967-1978.

Kim Y, Nielsen R. 2004. Linkage disequilibrium as a signature of selective sweeps. Genetics 167:1513-1524.

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, et al. 2002. A high-resolution recombination map of the human genome. Nat. Genet. 31:241-247.

Kopp M, Servedio MR, Mendelson TC, Safran RJ, Rodríguez RL, Hauber ME, Scordato EC, Symes LB, Balakrishnan CN, Zonana DM, et al. 2017. Mechanisms of Assortative Mating in Speciation with Gene Flow: Connecting Theory and Empirical Research. Am. Nat. 191:1-20.

Linn CE, Dambroski HR, Feder JL, Berlocher SH, Nojima S, Roelofs WL. 2004. Postzygotic isolating factor in sympatric speciation in Rhagoletis flies: reduced response of hybrids to parental host-fruit odors. Proc. Natl. Acad. Sci. U. S. A. 101:17753-17758.

Mallet J. 2005. Speciation in the 21st century. Heredity 95:105-109.

Mathyer ME, Brettmann EA, Schmidt AD, Goodwin ZA, Oh IY, Quiggle AM, Tycksen E, Ramakrishnan N, Matkovich SJ, Guttman-Yassky E, et al. 2021. Selective sweep for an enhancer involucrin allele identifies skin barrier adaptation out of Africa. Nat. Commun. 12:2557.

McVean GA, Charlesworth B. 2000. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. Genetics 155:929-944.

Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P, Feder JL. 2010. Widespread genomic divergence during sympatric speciation. Proc. Natl. Acad. Sci. 107:9724-9729.

Nam K, Munch K, Hobolth A, Dutheil JY, Veeramah KR, Woerner AE, Hammer MF, Project GAGD, Mailund T, Schierup MH, et al. 2015. Extreme selective sweeps independently targeted the

X chromosomes of the great apes. Proc. Natl. Acad. Sci. 112:6413-6418.

Nam K, Nhim S, Robin S, Bretaudeau A, Nègre N, d'Alençon E. 2020. Positive selection alone is sufficient for whole genome differentiation at the early stage of speciation process in the fall armyworm. BMC Evol. Biol. 20:152.

Pavlidis P, Živković D, Stamatakis A, Alachiotis N. 2013. SweeD: likelihood-based detection of selective sweeps in thousands of genomes. Mol. Biol. Evol. 30:2224-2234.

Schliewen UK, Tautz D, Pääbo S. 1994. Sympatric speciation suggested by monophyly of crater lake cichlids. Nature 368:629-632.

Servedio MR, Doorn GSV, Kopp M, Frame AM, Nosil P. 2011. Magic traits in speciation: 'magic' but not rare? Trends Ecol. Evol. 26:389-397.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. Genome Res. 17:520-526.

Via S. 2012. Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 367:451-460.

Via S, West J. 2008. The genetic mosaic suggests a new role for hitchhiking in ecological speciation. Mol. Ecol. 17:4334-4345.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A Map of Recent Positive Selection in the Human Genome. PLoS Biol 4:e72.

Weir BS, Cockerham CC. 1984. Estimating F-statistics for the analysis of population structure. Evolution 38:1358-1370.

Wu C-I. 2001. The genic view of the process of speciation. J. Evol. Biol. 14:851-865.