Research

Recombination and base composition: the case of the highly self-fertilizing plant *Arabidopsis thaliana*

G Marais*, B Charlesworth* and SI Wright*†

Addresses: *Institute of Cell, Animal and Population Biology, University of Edinburgh, EH9 3JT Edinburgh, UK. †Current address: Department of Biology, York University, 4700 Keele St, Toronto, Ontario M3J 1P3, Canada.

Correspondence: SI Wright. E-mail: stephenw@yorku.ca

Published: 14 June 2004

Genome ***Biology*** 2004, **5:**R45

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/7/R45

Received: 26 March 2004
Revised: 26 April 2004
Accepted: 30 April 2004

Abstract

Background: Rates of recombination can vary among genomic regions in eukaryotes, and this is believed to have major effects on their genome organization in terms of base composition, DNA repeat density, intron size, evolutionary rates and gene order. In highly self-fertilizing species such as *Arabidopsis thaliana*, however, heterozygosity is expected to be strongly reduced and recombination will be much less effective, so that its influence on genome organization should be greatly reduced.

Results: Here we investigated theoretically the joint effects of recombination and self-fertilization on base composition, and tested the predictions with genomic data from the complete *A. thaliana* genome. We show that, in this species, both codon-usage bias and GC content do not correlate with the local rates of crossing over, in agreement with our theoretical results.

Conclusions: We conclude that levels of inbreeding modulate the effect of recombination on base composition, and possibly other genomic features (for example, transposable element dynamics). We argue that inbreeding should be considered when interpreting patterns of molecular evolution.

Background

Recombination is probably a key factor in the evolution of genome organization in species such as yeast, mammals, *Drosophila* and *C. elegans*. In these species, genomic features such as nucleotide polymorphism [1-4], GC content [1,5-8], codon bias [6,9], intron size [10,11], transposable element density [12-14] substitution rates [15-17] and gene order [18] vary widely within the genome, and are correlated with the local rate of crossing over. These observations are often explained as the result of various processes such as selective sweeps, background selection and weak Hill-Robertson interference (wHR), which all cause a reduction in the efficacy of natural selection in regions of reduced crossing over [19-21].

Rates of crossing over have been shown to correlate not only with the GC content of synonymous sites, where weak natural selection is expected to act on codon-usage bias, but also with the GC content of noncoding sites [6,22]. This is unlikely to be because GC bases are recombinogenic, as the correlation is far stronger with silent DNA than with total DNA [8]; see also [23]. This unexpected correlation may reflect the action of weak selection on noncoding GC, which would be less effective in regions of reduced recombination [24]. Alternatively, it could be an effect of biased gene conversion [8,25,26]. Biased gene conversion (BGC) is a process that preferentially converts A/T into G/C at sites heterozygous for AT and GC. The net effect of BGC is to increase the GC content of

recombining DNA sequences. Assuming that the rate of this process is correlated with the rate of crossing over, BGC could therefore generate the observed increase in GC content in regions of high crossing over. An excess of AT→GC mutations in regions of high recombination could also lead to the observed correlation between GC content and recombination [27]. The relative importance of BGC, mutational biases, and wHR in driving these patterns remains unresolved [22,28], although BGC may be the most likely explanation, especially in organisms such as yeast and mammals, where there is a strong correlation between recombination and GC content [7].

To date, most analyses of the role of recombination in determining genome structure have been done on outcrossing species, with the notable exception of the presumably partial selfer *C. elegans* [6], whose selfing rate is not precisely known. In contrast, less attention has been given to *Arabidopsis thaliana*, which is known to be an almost complete selfer with a selfing rate of approximately 99% in the natural populations that have been studied [29,30]. High levels of inbreeding, as in *A. thaliana*, are expected to have important effects on the genomic structuring of base composition. Inbreeding leads to a strong increase in levels of homozygosity, which reduces the effective rate of recombination [31]. Therefore, processes sensitive to recombination and homozygosity, such as the effectiveness of selection on codon usage and the strength of BGC, will be affected by the high level of inbreeding apparently experienced by *A. thaliana* [7,32].

Previous work has provided evidence for a correlation between gene expression and codon bias in *Arabidopsis* [33,34], although the effect is weak. This suggests that translational selection is acting on codon bias in *Arabidopsis*. However, on the basis of the genes studied so far, no striking difference in codon bias between *A. thaliana* and its outcrossing congener *A. lyrata* has been observed, perhaps because of the population history of these species obscuring the expected patterns of molecular evolution [35]. Here we investigate the effect of inbreeding on the evolution of base composition (GC content and codon bias) within the *A. thaliana* genome, both theoretically and by DNA sequence analysis. Our goal was to test for an effect of recombination on GC content and codon bias in *A. thaliana*, and to use models to examine the joint effects of recombination and inbreeding on selection on codon usage and BGC, in order to help us to interpret the results from genome analysis. We show by computer simulation and modeling that selection on codon usage is not expected to vary with local recombination rate in a highly inbred species and that BGC is expected to be ineffective compared with an outcrossing species. We show that these predictions are consistent with the results of our analysis of the *A. thaliana* genome. We find no association between the local rate of crossing over and either codon bias or GC content (for both coding and noncoding regions).

Results

Recombination and codon usage

Previous simulation results have shown that, in outcrossing species, weak selection on codon usage is expected to be significantly reduced in genomic regions with low rates of crossing over because of wHR effects [21,36,37]. We modified the model of [21] by adding one additional parameter, the selfing rate S (see Materials and methods). The results for S = 0%, 50% and 99% for several values of the population recombination rate $4N_e r$ (where N_e is effective population size and r is the per base rate of recombination) and two values of the strength of selection $4N_e s$ (where s is the selection coefficient) are presented in Figure 1. The effect of crossing over on the efficacy of selection on codon usage decreases strongly with S. For S = 99%, which is probably close to the true value for *A. thaliana* [29,30], virtually no effect of crossing over is observed. This result reflects the strong reduction in the range of effective rates of recombination present in a selfing genome; high levels of homozygosity dramatically reduce the effective rate of recombination [31], and therefore a given difference in r between two genomic regions will produce much greater differences in effective recombination rates in an outcrosser than in a selfer. Therefore, the theory predicts very weak or no associations between selection on codon usage and the rate of crossing over in *A. thaliana*. Results with intermediate selfing rates of 50% show a similar effect of recombination to that with complete outcrossing, suggesting the presence of a threshold level of inbreeding that leads to uncoupling of recombination from codon bias evolution.

In *A. thaliana*, selection on codon usage seems to be relatively weak [34]. Thus, it is quite hard to identify the so-called 'optimal' codons, which are preferentially retained by translational selection. In other species such as *Drosophila* and *C. elegans*, optimal codons have been shown to correspond to the most abundant tRNAs in cells [38,39]. Using tRNA gene number as a proxy for tRNA expression, we redefined the list of optimal codons in *A. thaliana* as those corresponding to major tRNAs (S.I.W, C.B.K.Yau, M. Looseley, and B. Meyers, unpublished data). The frequency of these newly defined optimal codons (hereafter denoted F_{op}) is more strongly correlated with the level of gene expression than was the case in previous work (Spearman rank coefficient R_s = 0.26 with $p < 10^{-4}$). This is consistent with the idea that this new index better captures translational selection on codon bias than do previous ones.

We then compared selection on codon usage measured by F_{op} with the rate of crossing over estimated from the comparison of genetic and physical maps for each chromosome arm (see Materials and methods). Figure 2 shows that there is no relationship between these parameters, with F_{op} being equal to approximately 0.49 throughout the genome. R_s = -0.02 with $p < 10^{-4}$. Although the p value is highly significant, the correlation coefficient implies that only 0.04% of the variability in F_{op} is explained by the rate of crossing over, and the p value

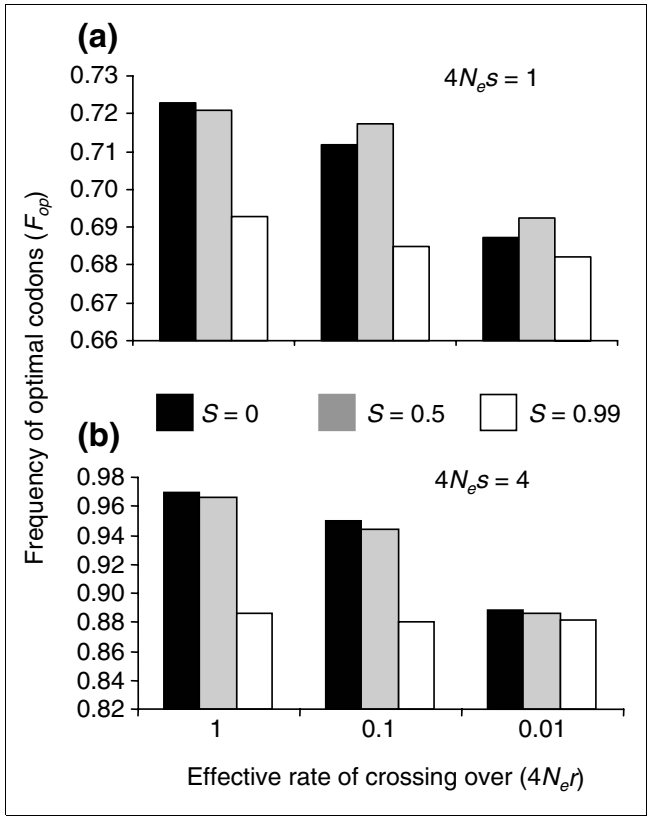

Figure 1
Simulation results showing the effect of recombination on the effectiveness of selection for codon usage under inbreeding. The figure shows the relationship between the recombination parameter $4N_er$ and the frequency of the optimal codon (F_{op}) for selfing rates (S) of 0 (black bars), 0.5 (gray bars) and 0.99 (white bars). **(a)** Value of the strength of selection $4N_es = 1$; **(b)** $4N_es = 4$. All simulations involved 1,000 sites where $4N_eu = 0.04$.

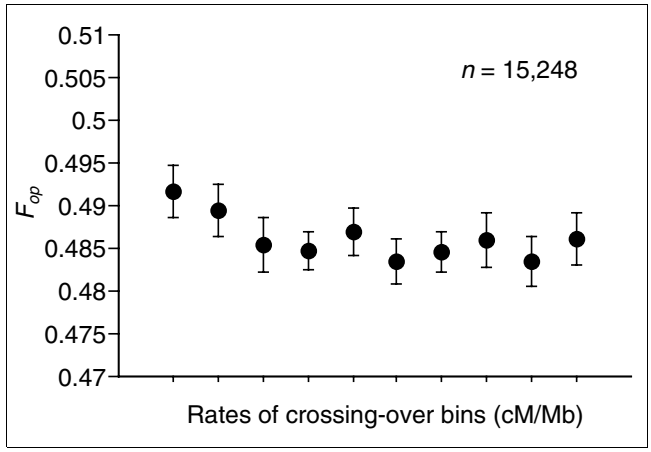

Figure 2
Genome-wide relationship between the frequency of optimal codons (F_{op}) and the rate of crossing over in *A. thaliana*. Each crossing-over bin contains approximately 10% of the genes of the dataset. See Materials and methods section for details of the measurement of F_{op} and the rate of crossing over, in centimorgans per megabase (cM/MB).

reflects the large number of genes used in the analysis. There is thus virtually no genome-wide correlation between F_{op} and the rate of crossing over. It has been noted previously for other species that the genome-wide correlation between codon bias and recombination is weak, although large differences can be observed among chromosomes or chromosomal regions with very different rates of crossing over [7]. This effect may result from poor map-based point estimates of recombination for any given locus, while global averages are much more reliable, as well as other causes [7,22]. In *A. thaliana*, as in many species, crossovers are suppressed near centromeres [40]. If we compare the centromeric regions with the remainder of the genome, we find at most a 1% difference: in centromeric regions, $F_{op} = 0.477$ ($n = 2005$), and in the other regions, $F_{op} = 0.486$ ($n = 13,243$). In contrast, comparisons of centromeric regions with other genomic regions in *Drosophila* show striking differences, which are larger than 20% [6]. Taken together, these observations suggest that selection on codon usage does not vary with the rate of crossing over in *A. thaliana*, in agreement with the theory.

Recombination and GC content

BGC can be seen as a sort of meiotic drive, in which GC gametes are favored over AT gametes [41]. As high levels of inbreeding are associated with a strong decrease in heterozygosity, the strength of BGC should be dramatically reduced in inbreeders, because BGC can occur only in heterozygotes. The expected change in GC content due to BGC in the case of inbreeding can be derived straightforwardly from population genetics theory (see Materials and methods for details). By using standard diffusion equations modifed for inbreeding one can obtain the GC content at equilibrium under BGC, mutation, drift and inbreeding (see Materials and methods for details). The GC content at equilibrium (p^*) depends on the effective population size (N_e), the mutational bias ($\alpha = u/v$, where u is the mutation rate from GC→AT and v the reverse mutation rate), the coefficient of BGC (ω), and the selfing rate (S) (see Materials and methods).

In Figure 3, we plot the expected values of p^* according to the scaled measure of the strength of BGC ($4N_e\omega$ with different mutational biases (α) and selfing rates (S)). In Figure 3a, we show that BGC has little effect on expected GC content in highly selfing populations ($S = 0.99$) compared to outcrossing populations ($S = 0$), regardless of the strength of BGC. This means that, in a highly selfing population, genomic regions with high recombination and thus high BGC (high ω) are expected to have a very similar GC content to genomic regions with low recombination and thus little or no BGC (low or null ω). Figure 3b shows that a slight difference between such genomic regions can be observed in a partial selfer, with $S = 50\%$, for example. In *A. thaliana*, where S has been estimated to be approximately 0.99 [29,30], the average GC contents for introns, 5' flanking regions and 3' flanking regions are 32.1%, 32.7% and 32.5%, respectively, with an overall mean of

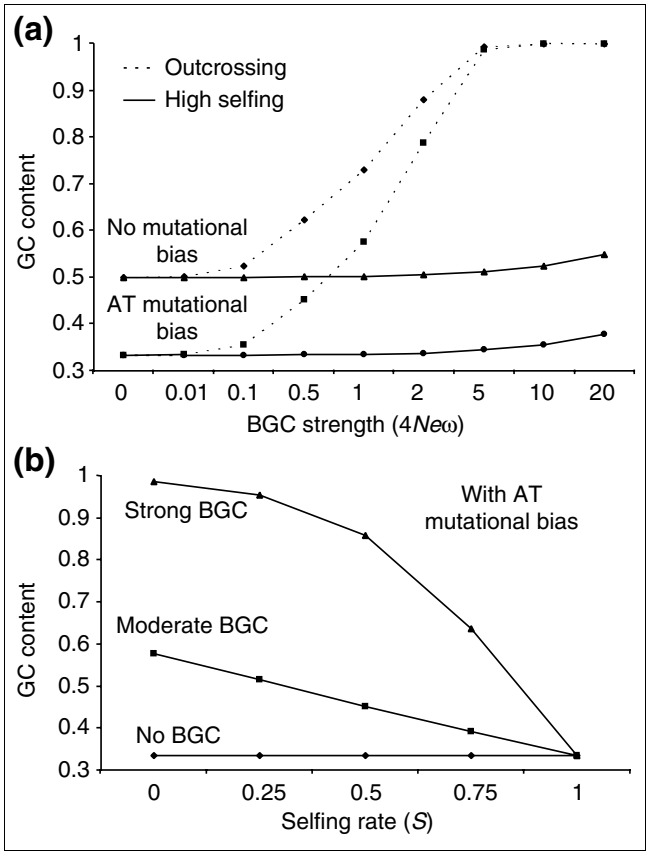

Figure 3
The effect of inbreeding on biased gene conversion (BGC). **(a)** The consequences of BGC for GC content for outcrossing populations (S = 0) and highly self-crossing populations (S = 0.99). The results are shown with no mutational bias (u/v = 1) or with a mutational bias towards AT (u/v = 2). **(b)** The GC content expected for various selfing rates with no ($4N_e\omega$ = 0), moderate ($4N_e\omega$ = 1) or strong ($4N_e\omega$ = 5) BGC. The mutational bias was set to 2. See Materials and methods for details of the model.

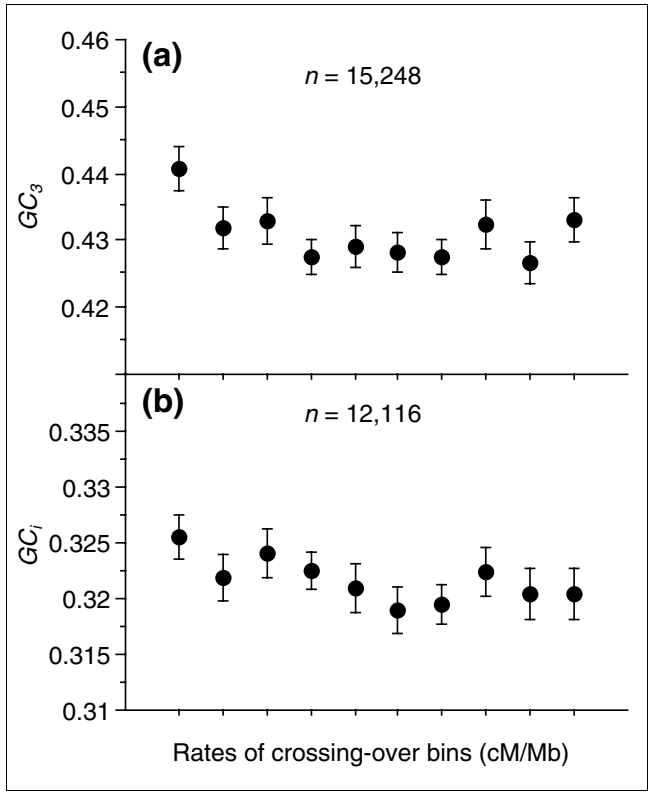

Figure 4
Genome-wide relationship between GC content and the rate of crossing over in *A. thaliana*. **(a)** Silent DNA (GC_3); **(b)** intron DNA (GC_i). Each crossing-over bin contains approximately 10% of the genes of the dataset. See Materials and methods for details of the measurement of GC content and the rate of crossing over, in centimorgans per megabase (cM/MB). The results are the same if only genes with total intron size greater than 100 nucleotides (or with larger cutoff values) are included.

approximately 32%. Thus, a mutational bias of 2 (that is, u/v = 2) describes well the average GC content of noncoding DNA (see Materials and methods for details). This suggests that the results obtained with S ~1 and α = 2, are probably the closest to reality in *A. thaliana*. With these parameters, Figure 3 shows that no effect of BGC on GC content is expected.

In Figure 4, we plot the GC content for third codon positions of coding DNA (GC_3) and intron DNA (GC_i) against the rate of crossing over in *A. thaliana*. No significant correlations between GC_3 and GC_i with recombination is observed. The correlation coefficients are very weak for both GC_3 (R_s = -0.03 with p = 0.0002) and GC_i (R_s = -0.04 with p = 0.01). In both cases, less than 0.2% of the variability in GC content is explained by recombination. Again, we checked for a difference between centromeric regions and the remainder of the genome. In centromeric regions, GC_3 = 41.4% and GC_i = 32.3%, and in the other regions GC_3 = 43.3% and GC_i = 32.8%. Thus, we observe a 2% difference for GC_3 and a 0.5% difference for GC_i. Again these differences have a p value lower than 0.05 (with a nonparametric Kolmogorov test) but these minor differences may have no biological meaning. In contrast, the corresponding difference in GC content in *Drosophila* is as large as 20% for coding and 5% for noncoding DNA [6]. Our results from the genome analysis thus seem to be in agreement with theory.

Discussion

Base composition in inbreeders versus outcrossers

Our genome analysis suggests that recombination has little effect on base composition in *A. thaliana*. Neither codon usage bias nor GC content are correlated with the local rate of crossing over. Our theoretical work suggests that, first, selection on codon usage is not expected to vary with crossing over in highly inbred species and, second, that BGC is inefficient in highly inbred species. We also expect the global efficacy of selection on codon usage to be lower in inbred than in outbred

species (see Figure 1). Interestingly, the level of codon usage is low in *A. thaliana* and high in *Drosophila* [34], in agreement with the respective levels of outcrossing in these species. Subsequent comparisons of selfing versus outcrossing *Arabidopsis* species have, however, shown a less clear pattern [35].

The budding yeast *Saccharomyces cerevisiae* is thought to have a high level of inbreeding in natural populations [42], and recent high estimates of the inbreeding coefficient from a population of the close relative *S. paradoxus* [43] suggest that long-term rates of selfing may be high. One might therefore expect a similar pattern in the genome of *S. cerevisiae* to that observed in *A. thaliana*. Although there is no evidence for a strong effect of recombination on the rate of protein evolution once gene expression is controlled for [44], recombination rates are strongly correlated with GC content in yeast [8]. However, in contrast to *A. thaliana*, yeast has exceptionally high rates of recombination, approximately 100-fold higher than the multicellular model systems *Arabidopsis*, *Drosophila* and *C. elegans* [7]. This may counteract the reduction in effective recombination rates caused by inbreeding, to such a degree that the strength of BGC may be significant. Indeed, a study of nucleotide variation at a prion-like gene in *S. cerevisiae* estimated a similar effective rate of recombination to that in *Drosophila* [45], despite possibly high rates of inbreeding in budding yeast.

In *C. elegans*, the level of codon bias is intermediate between that of *A. thaliana* and *Drosophila* [34], and there is also a significant correlation between crossing over and GC content in this species [6]. This is puzzling, in view of the low levels of genetic diversity (suggesting a low effective population size) [46,47] and the high levels of linkage disequilibrium (suggesting very restricted recombination due to inbreeding) [46]. There are three possible explanations for this pattern: first, *C. elegans* is in fact a fairly outcrossing species, and has recently suffered a population bottleneck that reduced its levels of genetic variability; second, it has only become a self-fertilizing hemaphrodite relatively recently (this possibility cannot be excluded, since we lack knowledge of its close relatives) [48]; and third, our models of the evolution of base composition are in error. Further information on the evolutionary biology of *C. elegans* and its relatives is needed to solve this problem.

How to explain variation in base composition

Base composition is fairly variable across the *A. thaliana* genome, both for codon bias and GC content (see Figure 5). Recombination does not seem to be a determinant of this variation in *A. thaliana*. What could be the other possible determinants? It is well known that codon bias has multiple determinants: gene expression and protein length, for instance [34]. Here, we find that a total of approximately 20% of the variability in F_{op} is explained by gene expression (measured by expressed sequence tag (EST) or massively parallel signature sequencing (MPSS) data) and protein length (S.I.W., C.B.K. Yau, M. Looseley, and B. Meyers, unpublished data), leaving 80% to be explained. The rate of nonsynonymous substitutions per site (d_N) seems to be another strong determinant of codon bias in *Drosophila* [16]. However, no large-scale dataset of orthologous pairs between *A. thaliana* and its close relatives (required to estimate d_N) is currently available, so we cannot assess the contribution of d_N to variability in F_{op}. Both the GC content at synonymous sites and at introns are likely to be influenced by genetic drift [49]. The cumulative effects of mutation, selection (in the case of synonymous sites) and drift should generate random variation in GC content across the genome. This can be explored by looking whether the distribution of GC content over genes (see Figure 5) follows a binomial distribution. However, the differences between the expected values (estimated using the mean GC content) and the observed values were statistically significant for both GC_3 and GC_i (data not shown). Variation in GC content does not seem to be fully explained by the effects of genetic drift.

This suggests that other factors, such as local differences in level of mutational bias, may contribute to the patterns of base composition in *A. thaliana*. If there are strong local effects driving the variation in GC, we would expect a strong positive correlation between GC_3 and GC_i across genes, as observed in humans [26,50] and *Drosophila* [51]. In contrast, we find a weak but significant negative correlation between GC_3 and GC_i (R_s = -0.115; $p < 10^{-4}$); this is the case even when the correlation between GC_3 and codon bias is factored out (R_s = -0.115; $p < 10^{-4}$). This suggests that local heterogeneity in mutational bias does not explain the variation in GC, and provides further evidence against a major effect of BGC in local GC variation in this species. The uncoupling of GC_3 and GC_i, and the residual variation in GC content, may result from the action of selective constraint on some intron sequences, and differences in the mutational context of introns and synonymous sites.

One important assumption of our analysis is that *A. thaliana* has been self-fertilizing for a sufficiently long time to remove any historical effects of recombination on base composition. This is questionable, given the fact that its closest known relatives are all obligate outcrosssers, including *A. lyrata* [52]. The most extreme case is complete cessation of outcrossing and complete relaxation of selection or BGC since the divergence of *A. thaliana* from *A. lyrata*. Under this assumption, Equation (4) given in Materials and methods implies that the present-day deviation of GC content of a genomic region of *A. thaliana* from the completely neutral value is equal to the initial deviation multiplied by $\exp(-(u + v)t)$, where t is the time since divergence. This can be related to the expected DNA sequence divergence at completely unconstrained sites (after a Poisson correction for multiple hits), which is equal to $4\alpha vt/(1 + \alpha)$ [49]. From the base composition of *A. thaliana* centromeric DNA (see Results), we can

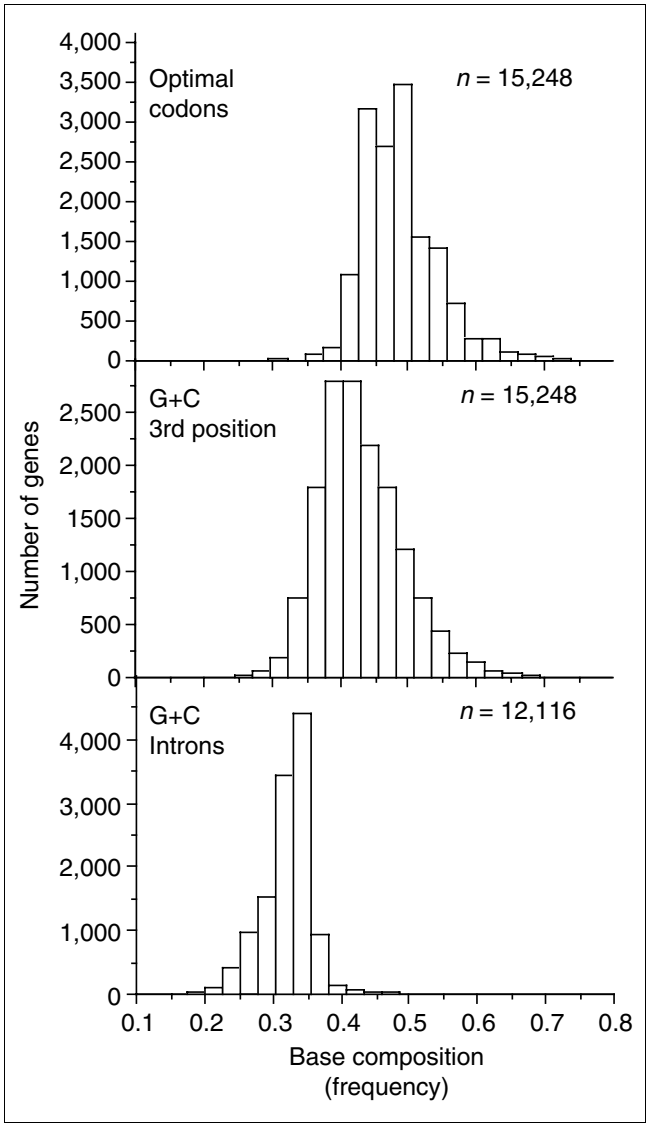

Figure 5
Distribution of base composition among genes in *A. thaliana*.

estimate α as 2.12, assuming that this reflects mutational equilibrium. Given that the maximum silent-site divergence between the two species is about 0.2 [35], this implies that $(u + v)t = 0.23$, so that the current deviation of *A. thaliana* GC content from the mutational equilibrium value is expected to be at least 80% of the value in *A. lyrata*. Conversely, this means that the maximal departure of GC content in a genomic region of *A. lyrata* from mutational equilibrium is at most a factor of $\exp((u + v)t)$ times that for the corresponding region of *A. thaliana*, that is, about 25% greater. If *A. thaliana* has only been highly selfing for a proportion of the time since divergence, the value will be proportionately lower. This suggests that regional variation in GC content in *A. lyrata* should be relatively modest, a prediction which can be tested when more genomic information on *A. lyrata* is available.

The influence of population subdivision

Our theoretical model of drift, selection and mutation considered only a single population, but it is known that *A. thaliana* has strong population structure [29,53]. In addition, within-population silent nucleotide site diversity is very low compared with *A. lyrata*, but the diversity when pooled over samples from different localities is similar for the two species [53]. This indicates that the effective population sizes of local populations are very low in *A. thaliana*, that migration among populations is limited, and that the effective population size determining the diversity among alleles randomly chosen from the species as a whole is not greatly reduced. The relatively high total diversity also suggests that local extinction and recolonization of populations does not play a major part in controlling genetic diversity [54].

This raises the question of what measure of effective population size is appropriate for determining the base composition under BGC or weak selection. In addition to mutational bias, theory shows that base composition is controlled by the relative fixation probabilities of new mutations from GC to AT and vice versa [21,55]. If migration is conservative (that is, the number of migrants entering each population equals the number leaving), with selection of the form of Equation (1) in Materials and methods, these fixation probabilities are controlled by the same N_e as is appropriate for the mean level of neutral diversity within demes, which is the same as for a population lacking any subdivision [56-58]. With nonconservative migration, rigorous theoretical results are not available, but heuristic models and computer simulations suggest that fixation probabilities will be usually be lower than with conservative migration [59-61]. These considerations imply that our conclusion, that fixation probabilities in *A. thaliana* will be closer to the neutral values than in *A. lyrata* for sites affected by BGC or weak selection, is either unaffected by population structure, or is a conservative one.

Conclusions

We have shown that inbreeding affects base composition by modulating the effectiveness of recombination. Inbreeding has also been shown to affect the dynamics of transposable elements [32,62-64]. Taken together, these studies suggest that mating system can have a major effect on genome organization, particularly when the levels of inbreeding are high, and should be taken into account when interpreting patterns of molecular evolution. Other population parameters such as demographic history [35,65,66] and population subdivision [64], should also be considered when analyzing patterns of genome evolution.

Materials and methods

Genomic approach

Sequence data

We wanted to build an ACNUC database for the *A. thaliana* complete genome, because this allows the user to make complex queries [67]. We required the *A. thaliana* complete genome to be in GenBank format to do this. As far as we know, the only release available in this format is release 1 (see [68]). However, the gene predictions in this release may contain some errors. To circumvent this problem, we used 15,248 genes for which we had evidence for gene expression (EST or MPSS, see below) and whose intron-exon structure has not changed from release 1 to release 3, increasing the chance that they correspond to true genes with accurate annotations. Coding sequences and intron sequences of these genes were used for further analysis.

Recombination data

We used the rates of crossing over from a previous study [64]. These were obtained by comparisons of genetic and physical maps for each chromosome arm. A polynomial was fitted to the data and the derivative of this polynomial curve was used to estimate the local rate of crossing over as a function of the position in the chromosome arm (for details and data see [64] and [69]). We could not use a sliding window approach to estimate the local rate of crossing over because of the scarcity of genetic markers.

Codon bias

This was estimated using the frequency of optimal codons (F_{op}). The list of optimal codons for *A. thaliana* was revised (S.I.W, C.B.K.Yau, M. Looseley, and B. Meyers, unpublished data) by identifying the optimal codons as those corresponding to the major tRNAs, whose cellular concentration was estimated from tRNA gene number following [39,70,71]. To check for a correlation between F_{op} and gene expression, we used EST data (as in [34]) and MPSS data (see [72]) as estimates of the level of gene expression. F_{op} was computed with a modified version of a previously described program [34].

Theoretical approach

Hill-Robertson interference and inbreeding

Computer simulations were run following the reversible mutation, selection and drift multilocus model of [20]. The model assumes equal rates of forward and back mutation, with a population mutation rate ($4N_eu$) of 0.04. Simulations were run assuming a scaled selection intensity $4N_es$ of 1 and 4, with 1,000 mutable sites, and a population size of $N_e = 100$. Although this effective size is likely to be an underestimate of the true value for *A. thaliana*, the most important determinants of the level of interference are the product of the scaled mutation parameter $4N_eu$ and the number of mutable sites, and the scaled selection coefficient $4N_es$ [20]. We modified the program to include the selfing rate S, where gametes are formed by random mating with probability $(1 - S)$, and by self-fertilization with probability S. As in [20], selection was additive, with codominant heterozygous effects at individual sites (that is, the relative fitness at an individual locus is $1 + s$ for the heterozygote, and 1 and $1 + 2s$ for the alternative homozygotes).

Biased gene conversion and inbreeding

BGC favours GC over AT in the context of recombination between polymorphic DNA sequences [7,8,26]. It is formally equivalent to meiotic drive, which acts only in heterozygtes to cause a departure from 1:1 Mendelian segregation of alleles [73]. Assume that alternative GC and AT alleles at a site are neutral and that the ratio of GC:AT gametes from GC/AT heterozygotes is $k:k - 1$. In a random-mating population, the change in frequency p after one generation of an allele with GC at a given site is [74]:

$$\Delta p = 2p(1 - p)(2k - 1) \quad (1)$$

If the population is inbred, the frequency of heterozygotes is reduced to $2p(1 - p)(1 - F)$, where F is the inbreeding coefficient [73]. At equilibrium under a mixture of selfing with probability S and random mating with probability $1 - S$, $F = S/(2 - S)$ [31]. After taking selfing into account, Equation (1) becomes:

$$\Delta p = \omega p\,(1 - p)(1 - F) \quad (2)$$

where $k = 0.5(1 + \omega)$.

We must also consider the effects of genetic drift on finite inbred populations. The effect of genetic drift in a single isolated population is inversely proportional to the effective population size (N_e). Under a wide range of conditions, N_e for an inbred population is approximately equivalent to that for a random mating population with otherwise similar demography, divided by $(1 + F)$ [31,75,76]. When BGC, mutation and genetic drift are weak, their effects are additive and we can work directly with the Li-Bulmer formula for equilibrium, which is derived from diffusion theory [21,49,55]. The GC content at equilibrium is given by the approximate equation:

$$p^* = \frac{\exp\left(4N_e\omega\left(1-S\right)\right)}{u/v+\exp\left(4N_e\omega\left(1-S\right)\right)} \quad (3)$$

where u is the rate of mutation from GC→AT and v is the reverse mutation rate.

If selection or BGC is completely relaxed after reaching an equilibrium (as the one given by Equation (3) for BGC), the process of change in GC content is described by the standard linear expression for change under mutation pressure [77]. The new equilibrium GC content, p^{**}, is equal to $v/(u + v)$, and the GC content at time t, p_t is given by:

$$p_t - p^{**} = (p^* - p^{**}) \exp(-(u + v)t). \quad (4)$$

Additional data files

Additional data available with this article online, show the codon bias and base composition in *Arabidopsis thaliana* (Additional data file 1). It lists all genes analyzed in our analysis of base composition, combined with point estimates of recombination rate and base composition for each gene.

Acknowledgements

We are grateful to Blake Meyers for sharing unpublished MPSS data and to Deborah Charlesworth for helpful comments on the manuscript. G.M. is a European Union Marie Curie postdoctoral fellow. B.C. is supported by a Royal Society Research Professorship, and S.W. was supported by a Commonwealth Fellowship and an NSERC postdoctoral fellowship.

References

1. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW, Weber JL: **Comparison of human genetic and sequence-based physical maps.** *Nature* 2001, **409:**951-953.
2. Lercher MJ, Hurst LD: **Human SNP variability and mutation rate are higher in regions of high recombination.** *Trends Genet* 2002, **18:**337-340.
3. Tenaillon MI, Sawkins MC, Anderson LK, Stack SM, Doebley J, Gaut BS: **Patterns of diversity and recombination along chromosome 1 of maize (*Zea mays* ssp. *mays* L.).** *Genetics* 2002, **162:**1401-1413.
4. Cutter AD, Payseur BA: **Selection at linked sites in the partial selfer *Caenorhabditis elegans*.** *Mol Biol Evol* 2003, **20:**665-673.
5. Eyre-Walker A, Hurst LD: **The evolution of isochores.** *Nat Rev Genet* 2001, **2:**549-555.
6. Marais G, Mouchiroud D, Duret L: **Does recombination improve selection on codon usage? Lessons from nematode and fly complete genomes.** *Proc Natl Acad Sci USA* 2001, **98:**5688-5692.
7. Marais G: **Biased gene conversion: implications for genome and sex evolution.** *Trends Genet* 2003, **19:**330-338.
8. Birdsell JA: **Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution.** *Mol Biol Evol* 2002, **19:**1181-1197.
9. Hey J, Kliman RM: **Interactions between natural selection, recombination and gene density in the genes of *Drosophila*.** *Genetics* 2002, **160:**595-608.
10. Carvalho AB, Clark AG: **Intron size and natural selection.** *Nature* 1999, **401:**344.
11. Comeron JM, Kreitman M: **The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces.** *Genetics* 2000, **156:**1175-1190.
12. Duret L, Marais G, Biemont C: **Transposons but not retrotransposons are located preferentially in regions of high recombination rate in *Caenorhabditis elegans*.** *Genetics* 2000, **156:**1661-1669.
13. Rizzon C, Marais G, Gouy M, Biemont C: **Recombination rate and the distribution of transposable elements in the *Drosophila melanogaster* genome.** *Genome Res* 2002, **12:**400-407.
14. Bartolome C, Maside X, Charlesworth B: **On the abundance and distribution of transposable elements in the genome of *Drosophila melanogaster*.** *Mol Biol Evol* 2002, **19:**926-937.
15. Williams EJ, Hurst LD: **The proteins of linked genes evolve at similar rates.** *Nature* 2000, **407:**900-903.
16. Betancourt AJ, Presgraves DC: **Linkage limits the power of natural selection in *Drosophila*.** *Proc Natl Acad Sci USA* 2002, **99:**13616-13620.
17. Hellmann I, Ebersberger I, Ptak SE, Paabo S, Przeworski M: **A neutral explanation for the correlation of diversity with recombination rates in humans.** *Am J Hum Genet* 2003, **72:**1527-1535.
18. Pal C, Hurst LD: **Evidence for co-evolution of gene order and recombination rate.** *Nat Genet* 2003, **33:**392-395.
19. Smith JM, Haigh J: **The hitch-hiking effect of a favourable gene.** *Genet Res* 1974, **23:**23-25.
20. Charlesworth B, Morgan MT, Charlesworth D: **The effect of deleterious mutations on neutral molecular variation.** *Genetics* 1993, **134:**1289-1303.
21. McVean GA, Charlesworth B: **The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation.** *Genetics* 2000, **155:**929-944.
22. Marais G, Mouchiroud D, Duret L: **Neutral effect of recombination on base composition in *Drosophila*.** *Genet Res* 2003, **81:**79-87.
23. Meunier J, Duret L: **Recombination drives the evolution of GC-content in the human genome.** *Mol Biol Evol* 2004, **21:**984-990.
24. Charlesworth B: **The effect of background selection against deleterious mutations on weakly selected linked variants.** *Genet Res* 1994, **63:**213-227.
25. Eyre-Walker A: **Recombination and mammalian genome evolution.** *Proc R Soc Lond B Biol Sci* 1993, **252:**237-243.
26. Galtier N, Piganeau G, Mouchiroud D, Duret L: **GC-content evolution in mammalian genomes: the biased gene conversion hypothesis.** *Genetics* 2001, **159:**907-911.
27. Perry J, Ashworth A: **Evolutionary rate of a gene affected by chromosomal position.** *Curr Biol* 1999, **9:**987-989.
28. Kliman RM, Hey J: **Hill-Robertson interference in *Drosophila melanogaster*: reply to Marais, Mouchiroud and Duret.** *Genet Res* 2003, **81:**89-90.
29. Abbott RJ, Gomes MF: **Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh.** *Heredity* 1989, **62:**411-418.
30. Berge G, Nordal I, Hestmark G: **The effect of breeding systems and pollination vectors on the genetic variation of small plant populations within an agricultural landscape.** *OIKOS* 1998, **81:**17-29.
31. Nordborg M: **Linkage disequilibrium gene trees and selfing: an ancestral recombination graph with partial self-fertilization.** *Genetics* 2000, **154:**923-929.
32. Charlesworth D, Wright SI: **Breeding systems and genome evolution.** *Curr Opin Genet Dev* 2001, **11:**685-690.
33. Chiapello H, Lisacek F, Caboche M, Henaut A: **Codon usage and gene function are related in sequences of *Arabidopsis thaliana*.** *Gene* 1998, **209:**GC1-GC38.
34. Duret L, Mouchiroud D: **Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*.** *Proc Natl Acad Sci USA* 1999, **96:**4482-4487.
35. Wright SI, Lauga B, Charlesworth D: **Rates and patterns of molecular evolution in inbred and outbred *Arabidopsis*.** *Mol Biol Evol* 2002, **19:**1407-1420.
36. Comeron JM, Kreitman M, Aguade M: **Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*.** *Genetics* 1999, **151:**239-249.
37. Tachida H: **Molecular evolution in a multisite nearly neutral mutation model.** *J Mol Evol* 2000, **50:**69-81.
38. Moriyama EN, Powell JR: **Codon usage bias and tRNA abundance in *Drosophila*.** *J Mol Evol* 1997, **45:**514-523.
39. Duret L: **tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes.** *Trends Genet* 2000, **16:**287-289.
40. Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408:**796-815.
41. Bengtsson BO: **Biased conversion as the primary function of recombination.** *Genet Res* 1986, **47:**77-80.
42. Fingerman EG, Dombrowski PG, Francis CA, Sniegowski PD: **Distribution and sequence analysis of a novel Ty3-like element in natural *Saccharomyces paradoxus* isolates.** *Yeast* 2003, **20:**761-70.
43. Johnson LJ, Koufopanou V, Goddard MR, Hetherington R, Schafer SM, Burt A: **Population genetics of the wild yeast *Saccharomyces paradoxus*.** *Genetics* 2004, **166:**43-52.
44. Pal C, Papp B, Hurst LD: **Does the recombination rate affect the efficiency of purifying selection? The yeast genome provides a partial answer.** *Mol Biol Evol* 2001, **18:**2323-2326.
45. Jensen MA, True HL, Chernoff YO, Lindquist S: **Molecular population genetics and evolution of a prion-like protein in *Saccharomyces cerevisiae*.** *Genetics* 2001, **159:**527-535.
46. Koch R, van Luenen HG, van der Horst M, Thijssen KL, Plasterk RH: **Single nucleotide polymorphisms in wild isolates of *Caenorhabditis elegans*.** *Genome Res* 2000, **10:**1690-1696.
47. Graustein A, Gaspar JM, Walters JR, Palopoli MF: **Levels of DNA polymorphism vary with mating system in the nematode genus *Caenorhabditis*.** *Genetics* 2002, **161:**99-107.

48. Felix MA: **Genomes: a helpful cousin for our favourite worm.** *Curr Biol* 2004, **14:**R75-R77.
49. Li WH: **Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons.** *J Mol Evol* 1987, **24:**337-345.
50. Bernardi G: **The compositional evolution of vertebrate genomes.** *Gene* 2000, **259:**31-43.
51. Akashi H, Kliman RM, Eyre-Walker A: **Mutation pressure, natural selection, and the evolution of base composition in *Drosophila*.** *Genetica* 1998, **102-103:**49-60.
52. Koch MA, Haubold B, Mitchell-Olds T: **Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in *Arabidopsis*, *Arabis*, and related genera (Brassicaceae).** *Mol Biol Evol* 2000, **17:**1483-1498.
53. Bergelson J, Stahl E, Dudek S, Kreitman M: **Genetic variation within and among populations of *Arabidopsis thaliana*.** *Genetics* 1998, **148:**1311-1323.
54. Pannell JR, Charlesworth B: **Effects of metapopulation processes on measures of genetic diversity.** *Philos Trans R Soc Lond B Biol Sci* 2000, **355:**1851-1864.
55. Bulmer M: **The selection-mutation-drift theory of synonymous codon usage.** *Genetics* 1991, **129:**897-907.
56. Maruyama T: **Some invariant properties of a geographically structured finite population: distribution of heterozygotes under irreversible mutation.** *Genet Res* 1972, **20:**141-149.
57. Nagylaki T: **Geographical invariance in population genetics.** *J Theor Biol* 1982, **99:**159-172.
58. Nagylaki T: **Fixation indices in subdivided populations.** *Genetics* 1998, **148:**1325-1332.
59. Cherry JL, Wakeley J: **A diffusion approximation for selection and drift in a subdivided population.** *Genetics* 2003, **163:**421-428.
60. Whitlock MC: **Fixation probability and time in subdivided populations.** *Genetics* 2003, **164:**767-779.
61. Roze D, Rousset F: **Selection and drift in subdivided populations: a straightforward method for deriving diffusion approximations and applications involving dominance selfing and local extinctions.** *Genetics* 2003, **165:**2153-2166.
62. Wright SI, Le QH, Schoen DJ, Bureau TE: **Population dynamics of an Ac-like transposable element in self- and cross-pollinating *Arabidopsis*.** *Genetics* 2001, **158:**1279-1288.
63. Morgan MT: **Transposable element number in mixed mating populations.** *Genet Res* 2001, **77:**261-275.
64. Wright SI, Agrawal N, Bureau TE: **Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*.** *Genome Res* 2003, **13:**1897-1903.
65. Andolfatto P, Przeworski M: **A genome-wide departure from the standard neutral model in natural populations of *Drosophila*.** *Genetics* 2000, **156:**257-268.
66. Wall JD, Andolfatto P, Przeworski M: **Testing models of selection and demography in *Drosophila simulans*.** *Genetics* 2002, **162:**203-216.
67. Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G: **ACNUC - a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage.** *Comput Appl Biosci* 1985, **1:**167-172.
68. **Entrez Genome.** [http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome]
69. **Supplementary data for Wright *et al.*: Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*.** [http://www.genome.org/cgi/content/full/13/8/1897/DC1]
70. Ikemura T: **Codon usage and tRNA content in unicellular and multicellular organisms.** *Mol Biol Evol* 1985, **2:**13-34.
71. Percudani R: **Restricted wobble rules for eukaryotic genomes.** *Trends Genet* 2001, **17:**133-135.
72. **Directory of MPSS data pages.** [http://mpss.udel.edu]
73. Hartl DL, Clark AG: *Principles of Population Genetics* 3rd edition. Sunderland, MA: Sinauer; 1997:542.
74. Nagylaki T: **Evolution of a finite population under gene conversion.** *Proc Natl Acad Sci USA* 1983, **80:**6278-6281.
75. Pollak E: **On the theory of partially inbreeding finite populations. I. Partial selfing.** *Genetics* 1987, **117:**353-360.
76. Laporte V, Charlesworth B: **Effective population size and population subdivision in demographically structured populations.** *Genetics* 2002, **162:**501-519.
77. Sueoka N: **On the genetic basis of variation and heterogeneity of DNA base composition.** *Proc Natl Acad Sci USA* 1962, **48:**582-592.