MOLECULAR ECOLOGY

Molecular Ecology (2014) **23**, 311–324 doi: 10.1111/mec.12600

Shared selective pressure and local genomic landscape lead to repeatable patterns of genomic divergence in sunflowers

SEBASTIEN RENAUT,* GREGORY L. OWENS* and LOREN H. RIESEBERG*†

*Department of Botany, Biodiversity Research Centre, University of British Columbia, Vancouver, BC, Canada V6T 1Z4,
†Department of Biology, Indiana University, 1001 East Third Street, Bloomington, IN 47405, USA

Abstract

The repeated evolution of traits in organisms facing similar environmental conditions is considered to be fundamental evidence for the role of natural selection in moulding phenotypes. Yet, aside from case studies of parallel evolution and its genetic basis, the repeatability of evolution at the level of the whole genome remains poorly characterized. Here, through the use of transcriptome sequencing, we examined genomic divergence for three pairs of sister species of sunflowers. Two of the pairs (*Helianthus petiolaris – H. debilis* and *H. annuus – H. argophyllus)* have diverged along a similar latitudinal gradient and presumably experienced similar selective pressure. In contrast, a third species pair (*H. exilis – H. bolanderi*) diverged along a longitudinal gradient. Analyses of divergence, as measured in terms of F_{ST}, indicated little repeatability across the three pairs of species for individual genetic markers (SNPs), modest repeatability at the level of individual genes and the highest repeatability when large regions of the genome were compared. As expected, higher repeatability was observed for the two species pairs that have diverged along a similar latitudinal gradient, with genes involved in flowering time among the most divergent genes. Genes showing extreme low or high differentiation were more similar than genes showing medium levels of divergence, implying that both purifying and divergent selection contributed to repeatable patterns of divergence. The location of a gene along the chromosome also predicted divergence levels, presumably because of shared heterogeneity in both recombination and mutation rates. In conclusion, repeated genome evolution appeared to result from both similar selective pressures and shared local genomic landscapes.

Keywords: adaptation, *Helianthus*, mutation rate, parallel evolution, recombination rate, RNA-seq, speciation, transcriptome

Received 30 October 2013; revision received 15 November 2013; accepted 18 November 2013

Introduction

In the *Origin*, Darwin argued that animals belonging to distinct lines of descent may assume a close external resemblance if they are adapted to similar environmental conditions (Darwin 1859). His perceptive observations have long been taken as evidence of evolution through natural selection (for example, see Hoekstra 2006; Arendt & Reznick 2008; Bernatchez *et al.* 2010; Elmer & Meyer 2011; Jones *et al.* 2012). Recently, increasing attention has focused on examining the extent to which similar phenotypes have evolved through the same genetic mechanisms (i.e. parallel evolution, Wood *et al.* 2005; Manceau *et al.* 2010; Elmer & Meyer 2011; Jones *et al.* 2012; Conte *et al.* 2012). These studies have yielded evidence that the evolution of phenotypic traits, and their genetic underpinnings can sometimes be surprisingly repeatable, owing the force of natural selection and shared demographic parameters such as population size, dispersal distance and mutation rates (Ralph & Coop 2010).

Correspondence: Sebastien Renaut, Fax: +1 604 822 6089;
E-mail: sebastien.renaut@gmail.com

In a recent literature survey, Conte and colleagues (Conte *et al.* 2012) estimated the probability of gene reuse in cases of parallel evolution. They estimated the mean probability of gene reuse during independent bouts of evolution to a similar phenotype to be at least 30%. One of the best examples of a single gene responsible for the repeated evolution of a similar phenotype is the case of the melanocortin-1 receptor (*mc1r*) gene. Hoekstra *et al.* (2006) identified a single nucleotide substitution in *mc1r* having a major role in the repeated evolution of lighter coats in mice inhabiting sand dunes in Florida. More importantly, the same gene has been implicated in the evolution of pale coloration in lizards (Rosenblum *et al.* 2004), birds (Theron *et al.* 2001) and various other mammals, including the evolution of colour polymorphism in wooly mammoths (Römpler *et al.* 2006). These studies represent a remarkable example of the independent use of the same genetic mechanism during adaptation, suggesting a bias in the fraction of the genome modulated by selection which, in turn, results in a limited number of solutions to reach an adaptive optimum. Repeated use of the same genes or mutations during independent phenotypic evolution therefore would (at least partly) reflect constraints on the availability of beneficial mutations. However, much of the work on genetic changes has focused on parallel phenotypic evolution (Conte *et al.* 2012), and less is known about the broader overall predictability or repeatability of genomic divergence during evolution.

In plants, many of the best examples of parallel phenotypic changes involve flower colour. For example, a common type of flower colour transition that occurs repeatedly in some plant groups is a shift from blue or purple, typically insect-pollinated flowers, to red, hummingbird-pollinated flowers (Streisfeld & Rausher 2009). In addition, parallel evolution from blue to red flowers can occur via changes in the same developmental/regulatory pathway (anthocyanin pathway), yet involve different genes (Smith & Rausher 2011). Other examples of repeated phenotypic evolution in plants involve adaptation to serpentine soils, which are characterized by high heavy-metal content and low calcium-to-magnesium ratios. Species adapted to these soils often possess common adaptive traits distinct from closely related species, such as xeromorphic foliage, reduced stature and a more developed root system (Brady *et al.* 2005; Turner *et al.* 2010).

Candidate gene studies have exemplified how natural selection can act preferentially on certain genes and therefore lead to repeatable patterns of genetic divergence. Yet the candidate gene approach suffers from an inevitable ascertainment bias, as putative targets of selection are defined a priori. Are repeated genetic changes ubiquitous and therefore reflect greater overarching properties of the genome or do they represent the low hanging fruit of evolutionary studies? In a recent review on parallel evolution in the 'genomic era', Elmer & Meyer (2011) argue that to identify the homologous and nonhomologous variation important in natural populations, it is imperative to move beyond approaches that limit exploration to laboratory populations and a priori genetic expectations. It is certainly preferable to start by identifying loci involved in adaptation in a unbiased manner and then ask whether the proportion of repeated genetic changes is greater than a neutral expectation (Nadeau & Jiggins 2010). One of the few examples of such an approach comes from the study of wild populations of *Drosophila melanogaster* that are adapted to latitudinal gradients in Australia and the United States (Turner *et al.* 2008). Here, the authors found that many regions of the genome show similar patterns of differentiation in the United States and Australia providing evidence that these are influenced by similar spatially varying selection on separate continents.

Studies of parallel evolution also suffer from a more general bias in that they focus on populations that have adapted to the same phenotypic optima and consequently, shared natural selective pressure is usually invoked as the main driver of adaptation. However, as pointed out by Losos (2011), even when populations adapt to different phenotypic optima, shared selective constraints may still lead to repeatable patterns of phenotypic change. At the genome level, variation in mutation rates or recombination rates may promote or constrain genome divergence and thereby lead to repeatable patterns of divergence (Lynch 2007). However, few studies have compared the repeatability of genome evolution among natural populations that have adapted to similar (parallel evolution) or different phenotypic optima. This is what we set out to do using wild species of sunflowers, to dissect how both natural selection but also genetic constraints can lead to repeatable patterns of divergence.

Here, we employed a transcriptome sequencing approach to study the repeatability of genomic divergence in three pairs of sunflower sister species (*Helianthus petiolaris – H. debilis, H. bolanderi – H. exilis* and *H. annuus – H. argophyllus*) that are thought to have diverged in conditions of reduced gene flow due to geographic isolation (Fig. 1, Rogers *et al.* 1982; Timme *et al.* 2007). Transcriptome scans are inherently restricted to coding regions, as opposed to a genome sequencing approach that targets both coding and noncoding regions. Yet, they do offer the advantage of surveying hundreds of thousands of coding mutations scattered throughout the genome, compared with analysing a limited number of candidate genes identified a priori.

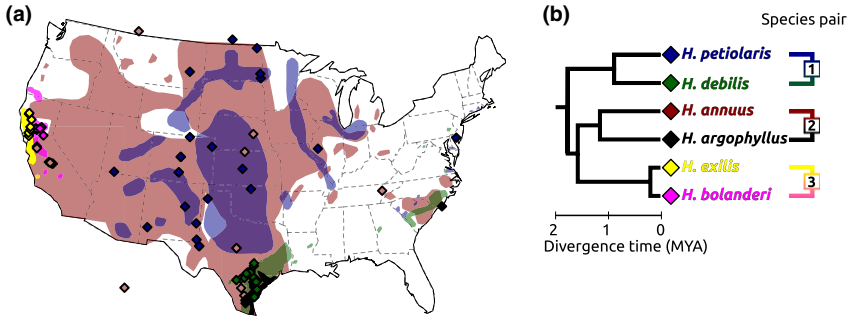

Fig. 1 Historical species range, sampling locations and phylogenetic relationship. (a) Species range (modified and redrawn from Rogers *et al.* 1982). For *Helianthus debilis*, the range shown is for two subspecies native to Texas because molecular phylogenetic analyses indicate that the other three subspecies (native to Florida) form a separate monophyletic taxon (Timme *et al.* 2007). (b) Phylogenetic relationship of three species pairs studied here.

The North American annual wild sunflowers are emerging as a model system for adaptation and speciation genomics research due to their well-characterized evolutionary history and adaptation to numerous environments (e.g. Rieseberg *et al.* 2003; Kane *et al.* 2011; Strasburg *et al.* 2011, 2012; Bowers *et al.* 2012; Andrew & Rieseberg 2013; Renaut *et al.* 2013a). The species are diploid (n = 17 chromosomes) annuals, with an obligate outcrossing mating system, which simplifies interpretation of population genomic data. *Helianthus annuus* is found throughout the central and western USA, while its closest relative *H. argophyllus* is endemic to southern Texas. Similarly, *Helianthus petiolaris* is widespread throughout the central and western USA, while its closest relative *H. debilis* is native to central and southern Texas. As such, these two species pairs have diverged along a similar latitudinal gradient from the great plains (*H. annuus* and *H. petiolaris*) to the southern United States (*H. argophyllus* and *H. debilis*) and appear to have adapted to similar changes in temperature, moisture and photoperiod regimes. The last species pairs we examined, *H. bolanderi* and *H. exilis*, have diverged along a different biogeographic gradient and share a more recent evolutionary history. *Helianthus exilis* is endemic to California's northern coast ranges, whereas *H. bolanderi* is mainly found in the central valley of California and in the eastern foothills of the Sierra Nevada mountain range.

The objectives of the study were to (i) assess the extent to which genomic divergence was repeatable; (ii) quantify how natural selection may affect repeatability and finally; and (iii) quantify the impact of local genome architecture on patterns of repeatability. Analyses of divergence indicated little repeatability at the level of individual single nucleotide polymorphism (SNP) markers, modest repeatability at the level of individual genes and the highest repeatability when large genomic regions were compared. In addition, we find higher repeatability when comparing the two species pairs that have diverged along a similar latitudinal gradient, indicative of the role of selection shaping genetic divergence. Lastly, repeatability in genomic divergence was also strongly influenced by selective constraints imposed, in part, by shared features of the local genomic landscape, especially shared heterogeneity in recombination rates and mutation rates among species.

Methods

Plant collection and transcriptome sequencing

Achenes (single-seeded fruits) representing 23 *Helianthus petiolaris*, 12 *H. debilis*, 21 *H. annuus*, 27 *H. argophyllus*, 8 *H. bolanderi* and 9 *H. exilis* spanning the range of each species were acquired either from USDA collections or from previous sampling efforts (Fig. 1, Table S1, Supporting information). Note that *H. petiolaris*, *H. debilis*, *H. annuus* and *H. argophyllus* individuals were sequenced as part of a previous study that identified genomic islands of divergence and tested predictions of divergence hitchhiking theory (Renaut *et al.* 2013a). All sequences from *H. bolanderi* and *H. exilis* individuals have not been reported previously. In addition, all species pairs comparisons and analyses presented here are distinct and novel from those reported in Renaut *et al.* (2013a) which focused on cases of divergence with gene flow.

Seeds were germinated at the University of British Columbia and grown for approximately 3 weeks in growth chambers (12 h of daylight at 22°) or similar greenhouse conditions, following which whole plants were harvested, flash frozen in liquid nitrogen and kept at −80°. For some plants, tissue from young undamaged leaves was collected directly in the field and flash frozen in liquid nitrogen. For each individual, RNA was extracted from young leaves tissue using a modified TRIzol Reagent (Invitrogen, Carlsbad, USA) protocol (Lai *et al.* 2012). Samples were quantified using a

NanoDrop (Thermo Fisher Scientific, Waltham, USA) and their quality verified on agarose gels. Total RNA was stored in pure water.

Approximately one-fifth of the samples (Table S1, Supporting information) were normalized, retrotranscribed and sequenced on an Illumina (San Diego, CA, USA) Genome Analyzer IIx at the David H. Murdock Research Institute (DHMRI, Kannapolis, NC, USA). cDNA was amplified and normalized with the TRIMMER-DIRECT cDNA Normalization Kit (Evrogen, Moscow, Russia). Other samples were sequenced on either an Illumina GAIIx at the Canada's Michael Smith Genome Sciences Center (GSC, Vancouver, Canada, www.bcgsc.ca/platform/solexa) or an Illumina HiSeq 2000 in the Biodiversity Research Center NextGen Sequencing Facility (BRC, Vancouver, Canada, sites.google.com/site/biodiversitynextgensequencing/home). Libraries were prepared following the standard Illumina Tru-Seq (LT) Protocol (pp. 35–69) with one slight modification. The RNA was not fragmented during the poly-A mRNA purification step, but directly reverse transcribed into cDNA. Upon cDNA purification, samples were sheared to approximately 400 bp on a Covaris (Woburn, MS, USA) sonicator. All individual libraries were uniquely barcoded prior to sequencing or run on a separate plate (paired end sequencing, 2 × 100 bp reads). Base calling was performed via the standard Illumina CASAVA pipeline.

Sequence alignments against reference

Reads were aligned against a reference transcriptome using the Burrows–Wheeler Aligner (BWA, ALN and SAMPE commands, Li & Durbin 2009). The reference data set consisting of 51 468 contigs (51.3M base pairs) is available on DRYAD (Renaut *et al.* 2013b) and described in a previous publication (Renaut *et al.* 2013a). Briefly, it was generated by sequencing four *H. annuus* libraries prepared from plants grown under different conditions and assembled using TRINITY (Grabherr *et al.* 2011).

Phylogenetic network

As relationships among populations may not conform to a tree-like pattern due to potential gene flow and shared ancestral polymorphisms, we performed a phylogenetic network analysis using the neighbour-net method implemented in SPLITSTREE4 (Huson 2005). We used SAMTOOLS (MPILEUP and BCFTOOLS, Li *et al.* 2009) to call single nucleotide polymorphisms (SNPs) using information from all samples for a subset of 1000 randomly selected genes. From this, we compiled an artificial nucleotide sequence comprising about 12 000 high-quality (overall missing data <10%) SNPs, coded according to the International Union of Pure and Applied Chemistry nucleotide code. We then used these markers to generate a phylogenetic network in SPLITSTREE4 (Huson 2005) using default parameters (Uncorrected_P distance as metric, ambiguous states ignored and normalize option accounting for unequal distribution of missing data across individuals, Fig. 2).

Independent variant calling per species pair

To generate three independent SNP data sets, we used SAMTOOLS (MPILEUP and BCFTOOLS, Li *et al.* 2009) to call single nucleotide polymorphisms (SNPs) for each species pair separately. Questionable SNPs were removed due to

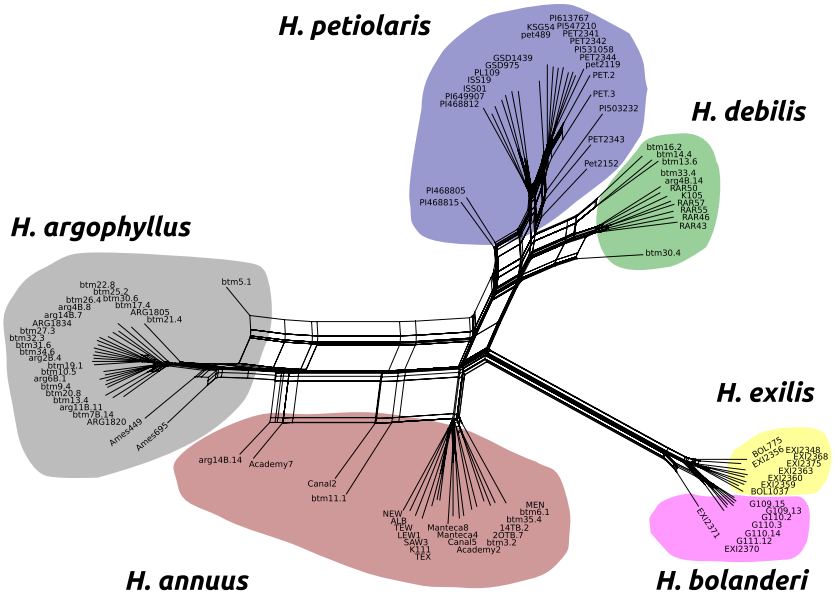

Fig. 2 Phylogenetic network reconstructed with the neighbour-net method. The network is based on 12 000 high quality SNPs (missing value < 10%) genotyped for all 100 individuals.

poor quality sequence, low coverage, potential sequencing errors and paralogy. Interspecific comparisons varied in terms of the number of individuals available per comparison and the sequence depth per individual. Therefore, we used different missing data thresholds (< 20% missing data for *H. petiolaris – H. debilis* and *H. annuus – H. argophyllus* and < 30% for *H. bolanderi – H. exilis*) so that the number of SNPs per comparison could be held roughly constant and sampling biases could be avoided (Table 1). We filtered out SNPs with low expected heterozygosity ($H_E < 0.2$), given that they either represent sequencing errors (unless very high coverage was attained) or rare alleles with little information content for interspecific comparisons. We also filtered out SNPs with very high observed heterozygosity ($H_O > 0.6$), because they likely represent paralogous sequence variants. Nevertheless, we recognize that our final data set likely contains a small fraction of false positives due to alignment and/or sequencing errors. Yet, given the large amount of data, high overall coverage, strict quality threshold cut-offs and visual inspection of random subsets of alignments (approximately 20 kb of alignments), we expect the data to be more than sufficient for the genome-wide analysis conducted here. From this curated SNP data set, F_{ST} values (Weir & Cockerham 1984) were calculated for each marker and each species pair, using the package HIERFSTAT (Goudet 2005) in the programming language R (R Core Team 2012).

Repeatable patterns of divergence: genome-wide approach

Next, we analysed the repeatability of patterns of divergence at three levels of genome organization: at the level of the mutation itself (SNP), at the level of the gene and at the level of large genomic regions. We tested for repeatable patterns of divergence for these three different comparisons (*H. petiolaris – H. debilis* versus *H. bolanderi – H. exilis*, *H. bolanderi – H. exilis* versus *H. annuus – H. argophyllus* and *H. annuus – H. argophyllus* versus *H. petiolaris – H. debilis*).

First, at the level of individual SNP markers, for each of the three comparisons aforementioned, we calculated Pearson's correlation coefficients between F_{ST} values for polymorphic SNP markers (*i.e.* SNPs that were polymorphic in both species pairs employed in each of the three comparisons). At the gene level, mean F_{ST} per gene was recorded as the average of all SNPs with $F_{ST} > 0$ within a gene, and correlation coefficients were calculated for polymorphic genes, for each of the three comparisons. Lastly, at the level of the whole genome, we used a previously published map to position (BLASTN) nearly half of all transcriptome contigs (24 406) onto 3047 unique genomic map locations covering all 17 *H. annuus* chromosomes (genetic map construction detailed in Renaut *et al.* 2013a). For each unique genomic position, mean F_{ST} was recorded by averaging all SNP markers found at that position (±0.001 cM). Then, we calculated correlation coefficients for genomic positions that were polymorphic for each of the three comparisons. Note that nonoverlapping genomic positions (windows) were used to avoid falsely increasing correlation values due to nonindependence of windows.

Repeatable patterns of divergence: quantile approach

We quantified patterns of repeatability with respect to level of divergence (F_{ST}). For this purpose, genes were ranked according to F_{ST} for each species pair and split into twenty groups (quantiles), each representing 5% of the ranked data. Then, the number of shared genes per quantile (for each of the three comparisons described above) was compared to the expected number of shared genes. For example, we identified one hundred and thirty-three genes that were in the first (top 5%) quantile both in the *H. petiolaris – H. debilis* and the *H. annuus – H. argophyllus* comparison, and this compares to a null expectation of thirty-three genes (5% * 5% * total number of polymorphic genes = 0.05 * 0.05 * 13 244 = 33). We repeated this analysis for each of the twenty quantiles, each representing 5% of all polymorphic genes for each of the three possible comparisons.

Table 1 Summary statistics

	Helianthus petiolaris – H. debilis	*H. bolanderi – H. exilis*	*H. annuus – H. argophyllus*
Number of SNPs	203 426	175 598	201 129
Mean F_{ST}	0.23	0.17	0.37
Number of genes with >2 reads aligned	28 046	26 063	29 696
Mean [median] number of reads aligned for genes with >2 reads aligned	757 [105]	944 [168]	616 [89]
Number of genes with >1 SNP (mean gene length = 998 bp)	17 395	19 140	16 612
Mean number of SNPs per gene [95% CI]	9.5 [9.5–9.7]	6.3 [6.2–6.3]	11.0 [10.9–11.2]
Number of polymorphic genomic regions (total = 3047 unique regions)	2781	2838	2775
Mean number of SNPs per genomic region [95% CI]	64 [59–70]	55 [51–60]	65 [60–70]

Finally, the test statistics lnRH (Schlotterer & Dieringer 2005) was used to determine whether genetic variance (in terms of observed heterozygosity) changed in the same direction for the two species pairs that have diverged along a similar latitudinal gradient. In other words, this allowed testing if markers showing elevated F_{ST} showed reductions in variance in the species found in the south (*H. debilis*/*H. argophyllus*) or the north (*H. petiolaris*/*H. annuus*).

Based on the results of the quantile analysis (Fig. 4), we used ErmineJ (Lee *et al.* 2005) to test for overrepresentation of gene ontology categories in the most and least divergent genes that were shared in the *H. petiolaris – H. debilis* versus *H. annuus – H. argophyllus* comparison. These are the species pairs that diverged along a similar latitudinal gradient and for which we might therefore expect certain traits, especially related to flower phenology and life history (Blackman *et al.* 2011; Kawakami *et al.* 2011), to experience similar selective pressures. We looked for overrepresentation of gene ontology terms in the most/least divergent shared genes (top/bottom 15%) compared to all other genes expressed. We arbitrarily chose a 15% cut-off to include quantiles (i) that appear to have diverged most strongly from expectations in terms of shared genes (see Fig. 4) and (ii) enough genes to provide adequate statistical power.

Properties of the genome influencing repeatable divergence

To identify properties of the genome that may explain the levels of repeatability, we tested how recombination rates (previous reported in Renaut *et al.* 2013a) influenced patterns of divergence (F_{ST}). We also calculated the synonymous substitution rate (d_s) as a proxy for mutation rate, under the assumption that silent sites are free from selective pressures (Eyre-Walker & Keightley 1999). To calculate d_s, we used the approach described in Renaut *et al.* (2012). Briefly, we identified the longest open reading frame, applied a majority rule for polymorphic sites to call consensus sequences for each species and calculated d_s using PAML (Yang 2007). We calculated the relationship between d_s and F_{ST} for each of the species pairs and then the relationship of d_s among species pairs.

Results

Alignments

We sequenced nearly 2 billion (10e9) paired-end reads (Table S1, Supporting information). Sequences were aligned against a reference transcriptome of 51 468 contiguous expressed sequences (contigs). On average, 56% of the raw reads aligned to the reference (Table S1, Supporting information) and all six species had a similar percentage of reads aligned, except for *H. exilis*, which had slightly fewer reads aligned than the other five species (48%, one-way ANOVA: $F_{5,94} = 4.5$, P-value = 0.001). The average number of fragments (one paired-end read equals two fragments) aligned was 21.4 million per individual (95% CI = 19.1–23.6). The percentage of reads aligned was mainly a result of RNA and sequencing library quality rather than species identity.

Phylogenetic network

We called polymorphic markers for a subset of 1000 random genes and assembled a set of 12 000 high quality SNPs genotyped for all 100 individuals (<10% missing data per marker). This allowed us to build a phylogenetic network to confirm the species identity of individuals, as well as to estimate relationships among species (Fig. 2). We observed that each species formed a distinct phylogenetic cluster as anticipated and previously reported (Renaut *et al.* 2013a; except for *H. exilis* and *H. bolanderi*, which had not been reported). However, a few individuals (namely btm5-1, arg14B-14, Academy7, which appear between the *H. annuus* and *H. argophyllus* clusters in Fig. 2) generate a conflicting signal that we suspect is due to significant levels of introgression. These individuals derive from contact zones in the southern distribution of *H. annuus*, so again this is not completely unexpected. Given their more recent divergence, *H. exilis* and *H. bolanderi* were the least differentiated species (Fig. 2).

Patterns of variability

We called variable sites for all reference genes and for each species pair to generate three independent SNP data sets. Based on strict quality thresholds, 203 427, 175 599 and 201 300 SNPs were detected for *H. petiolaris – H. debilis*, *H. bolanderi – H. exilis* and *H. annuus – H. argophyllus*, respectively (Table 1), with an average of 3.0 SNPs per 100 base pairs. From this curated data set, genetic divergence (F_{ST}) was calculated for each SNP. Mean F_{ST} for each species pairs was 0.23, 0.17 and 0.37 for *H. petiolaris – H. debilis*, *H. bolanderi – H. exilis* and *H. annuus – H. argophyllus*, respectively (Table 1).

Repeatability in patterns of variability

We calculated Pearson's correlation coefficients for individual SNPs for each of the three comparisons. All correlation coefficients were significant (P-value < 2e-16) but varied among comparisons, ranging from 0.06 for

H. petiolaris – H. debilis versus *H. bolanderi – H. exilis*, to 0.07 for *H. bolanderi – H. exilis* versus *H. annuus – H. argophyllus*, and 0.16 for the *H. annuus – H. argophyllus* versus *H. petiolaris – H. debilis* comparison (Table 2). In addition, it is important to note that these correlations are calculated from SNPs that were shared (*i.e.* polymorphic) in both comparisons tested. Including SNPs that were polymorphic in only one of the two species pairs would result in slightly negative correlations because a large proportion of markers that are variable in one species pair (*i.e.* $F_{ST} > 0$) are not variable in the other (*i.e.* $F_{ST} = 0$).

We then calculated mean F_{ST} per gene and following this, assessed the degree of correlation among species pairs. We observed higher correlations, ranging from 0.16 for *H. petiolaris – H. debilis* versus *H. bolanderi – H. exilis*, to 0.15 for *H. bolanderi – H. exilis* versus *H. annuus – H. argophyllus* to 0.33 for the *H. annuus – H. argophyllus* versus *H. petiolaris – H. debilis* comparison (Table 2, *P*-values < 2e-16). Finally, correlations in divergence at the genomic region level were the highest, with values of 0.24, 0.17 and 0.38 for the same three comparisons (Table 2 and Fig. 3a-c, *P*-values < 2e-16). Correlation coefficients also appear to be highly variable among chromosomes ranging from 0 to 0.6 (Fig. 3d).

To determine whether the higher correlation coefficients observed in the genomic regions analysis were a statistical artefact of averaging many markers over fewer regions (there were 3027 unique genomic locations compared with about 200 000 SNPs per comparison), we permuted F_{ST} for each SNP in each species pairs and recalculated mean F_{ST} per species pairs and correlation coefficients between genomic regions. We performed 1000 permutations for each of the three comparisons. Average correlation coefficients (1000 permutations × three comparisons) were not significantly different from zero (*t*-test, *P*-value = 0.48).

Repeatability in patterns of variability: quantile approach

We quantified patterns of repeatability with respects to level of divergence (F_{ST}) to test whether selection acts preferentially on certain genes in different comparisons. We observed that genes showing either extreme high or extreme low divergence tended to be shared among groups more than genes showing medium levels of divergence (Fig. 4, see also Fig. S1, Supporting information). Significant quadratic relationships between the twenty quantiles and the ratio of observed versus expected shared genes per quantile were identified (Fig. 4, *P*-values = 0.001, 0.0003, 2e-05 for *H. petiolaris – H. debilis* vs. *H. bolanderi – H. exilis*, *H. bolanderi – H. exilis* vs. *H. annuus – H. argophyllus* and *H. annuus – H. argophyllus* vs. *H. petiolaris – H. debilis*, respectively). This relationship was especially strong for the *H. petiolaris – H. debilis* vs. *H. annuus – H. argophyllus* comparison, in which the species pairs have diverged along a similar latitudinal gradient (Fig. 4, open circles). We also confirmed that the relationship was not due to difference in window size for each quantile [*i.e.* given the distribution of values, the range of F_{ST} values in the top quantiles are larger than for the middle quantiles where most of the data are distributed, Fig. S2 (Supporting information)].

While markers showing elevated F_{ST} values were shared between *H. annuus – H. argophyllus* and *H. petiolaris – H. debilis*, based on lnRH statistics, these do not appear to show similar directionality in reductions in heterozygosities (Fig. 5). For the *H. annuus – H. argophyllus* comparison, the top 5% most divergent markers show positive lnRH values (mean = 0.8, Fig. 5), implying reductions in heterozygosity for highly divergent markers in the southern species (*H. argophyllus*); while for *H. petiolaris – H. debilis*, the top 5% most divergent markers exhibit negative lnRH values (mean = −0.97, Fig. 5), which imply reductions in heterozygosity for highly divergent markers in the northern species (*H. petiolaris*).

Repeatability in patterns of variability: gene ontology

Three gene ontology (GO) terms (biological process) were overrepresented (corrected *P*-values < 0.05) among the list of the most/least (top/bottom 15%) divergent shared genes in the *H. annuus – H. argophyllus* vs. *H. petiolaris – H. debilis* comparison. Notably, the related

Table 2 Correlation coefficients at three different levels of genome organization. At the gene level, F_{ST} for each gene was calculated as the average of all SNPs with $F_{ST} > 0$ within a gene. Similarly at the level of the genome, all SNP markers found at a genetic map position were used to calculate average F_{ST} before correlations were calculated. All correlation coefficient significant (*P*-values < 2e-16).

		Pearson's correlation coefficient		
Species pairs 1	Species pairs 2	SNP	Gene	Genome
Helianthus petiolaris – H. debilis	*H. bolanderi – H. exilis*	0.06	0.16	0.24
H. annuus – H. argophyllus	*H. bolanderi – H. exilis*	0.07	0.15	0.17
H. annuus – H. argophyllus	*H. petiolaris – H. debilis*	0.18	0.33	0.38

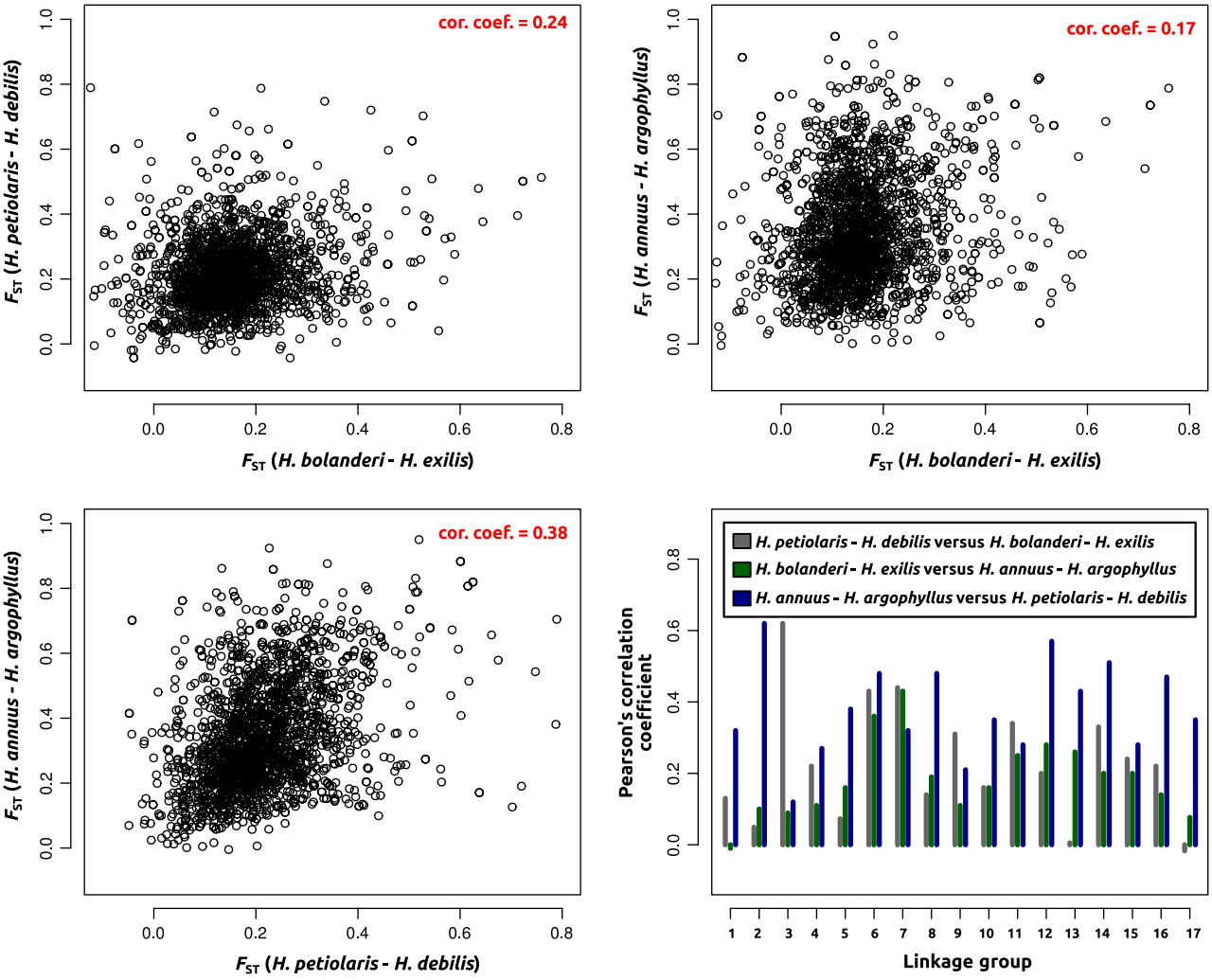

Fig. 3 Patterns of genetic divergence among species pairs. (a, top-left; b, top-right; c, bottom-left) Pearson's correlation coefficients calculated from averaging F_{ST} per genomic region for each species pair independently (d, bottom-right) Pearson's correlation coefficients calculated per chromosome for each of the three comparisons in (a–c).

GO categories *regulation of long-day photoperiodism, flowering* and *regulation of photoperiodism, flowering* were overrepresented in the high divergence shared gene list while *maintenance of inflorescence meristem identity* was the category overrepresented in the low divergence shared genes list (Table 3).

Repeatability in patterns of differentiation: causal factors

We then examined potential causes for the correlated patterns of divergence. We first examined the effects of recombination rate, as reductions in recombination are expected to increase genetic divergence (Nachman & Payseur 2012; Roesti *et al.* 2012; Renaut *et al.* 2013a), and therefore, if recombination rates are conserved among species, this should lead to correlated patterns of divergence. We found that recombination rates were negatively correlated with overall divergence (F_{ST} calculated per genomic region as described in methods) in all three comparisons (Fig. S3, Supporting information), as previously reported for other species pairs of wild sunflowers (Renaut *et al.* 2013a). In other words, regions of the genome having fewer genetic recombination events per unit of physical distance tended to be more divergent (Fig. S3, Supporting information).

We also calculated the rate of synonymous mutation per synonymous site (d_s), as a proxy for mutation rate (Baer *et al.* 2007). We observed that d_s was heterogeneous along the genome, ranging from 0 to 0.11 and that it was positively correlated with F_{ST} in all three species pairs (Fig. S4, Supporting information). In addition, d_s was also correlated among species pairs

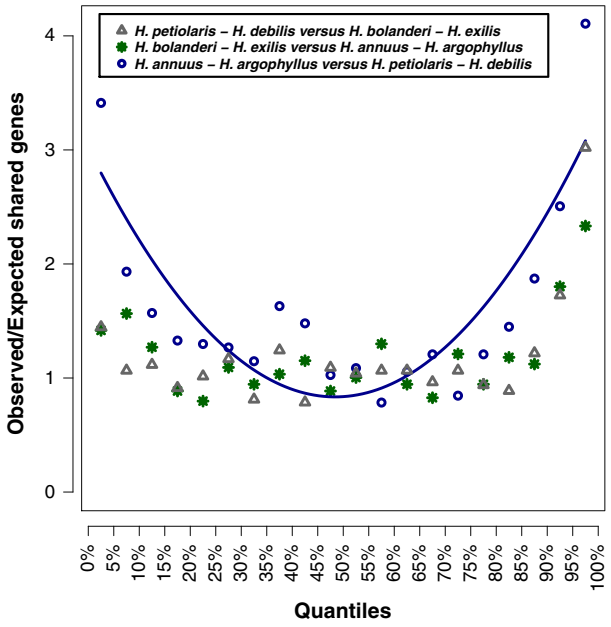

Fig. 4 Quantile (each representing 5% of the F_{ST} distribution) approach to patterns of repeatability. Patterns of repeatability quantified with respect to level of divergence (F_{ST}). Genes showing extreme high or extreme low divergence were more often shared among groups than genes showing medium levels of divergence. Significant quadratic relationships were observed between the twenty quantiles and the ratio of observed versus expected shared genes per quantile (quadratic curve for *Helianthus annuus – H. argophyllus* versus *H. petiolaris – H. debilis*, open circles).

[Fig. S5 (Supporting information), 0.17 to 0.19 to 0.23 for *H. petiolaris – H. debilis* vs. *H. bolanderi – H. exilis*, *H. bolanderi – H. exilis* vs. *H. annuus – H. argophyllus* and *H. annuus – H. argophyllus* vs. *H. petiolaris – H. debilis*, respectively], but differences in correlation coefficients did not vary significantly among comparisons (overlapping 95% CI).

Discussion

Genomic scans permit an unbiased view of the genetic changes underlying adaptive events, yielding insights into the predictability of evolution (Stern & Orgogozo 2009). In the present study, we observed evidence of repeatable patterns of divergence in three independent species pairs and at all levels of genomic organization, but especially at the level of genes and genomic regions. Our transcriptome scans support the view that genome divergence is largely repeatable and this appears to be caused both by adaptive and nonadaptive evolutionary forces. We find that both selection and the architecture of the genome influence patterns of repeatability. Below, we discuss the relative importance of each mechanism.

Effects of selection on patterns of repeatability

Two of the species pairs targeted by this study (*H. annuus – H. argophyllus* and *H. debilis – H. petiolaris*, Fig. 1) have diverged along a similar latitudinal cline (from the central great plains of the United States to southern Texas). As such, they likely have experienced a similar gradient of photoperiod, temperature and moisture regime and therefore overall similar selective pressure. It is well known that widely distributed plant species often exhibit latitudinal or altitudinal patterns of variation related to flowering phenology and several other life-history traits (Olsson & Agren 2002; Kawakami *et al.* 2011). For example, in perennial sunflowers (*H. maximiliani*), strong latitudinal differentiation was found for days to flowering, growth rate and multiple size-related traits and differentiation for these traits is likely a result of local adaptation driven by spatially heterogeneous environments (Kawakami *et al.* 2011). Similarly, *Helianthus annuus* populations across latitudinal transects are locally adapted to the particular photoperiod conditions encountered (Blackman *et al.* 2011). Based on these previous observations, we predicted and confirmed that patterns of repeatability were the highest between the two species pairs that have diverged along the same latitudinal cline. More importantly, gene ontology analyses also confirmed that biological processes related to flowering time were overrepresented among the highly divergent shared genes in the *H. annnus – H. argophyllus* versus *H. petiolaris – H. debilis* comparison (Table 3).

We observed relatively little repeatability at the level of the mutations (SNPs) themselves. Given that linkage disequilibrium is expected to decay within a few hundred base pairs to negligible levels (at least for *H. annuus*, Liu & Burke 2006), we can expect most markers to be independent from one another. As mutations within a gene arise randomly, it is relatively unlikely that the same mutation would appear twice in independent diverging lineages. In fact, only about 2% of all SNPs were found to be polymorphic in all three species pairs. These rare SNPs may in fact represent ancient standing genetic variation segregating within these species and which arose prior to the lineages splitting into separate species.

However, these fairly low levels of repeatability at the level of the mutation (SNP) contrast with higher repeatability at the level of the gene and genome. Correlation coefficients were about twice as high at the gene level than at the level of individual SNP. As expected, repeatability was higher for the comparison of divergence along the same latitudinal gradient, affirming the importance of natural selection (Table 2). These results imply that similar selection pressures among independent

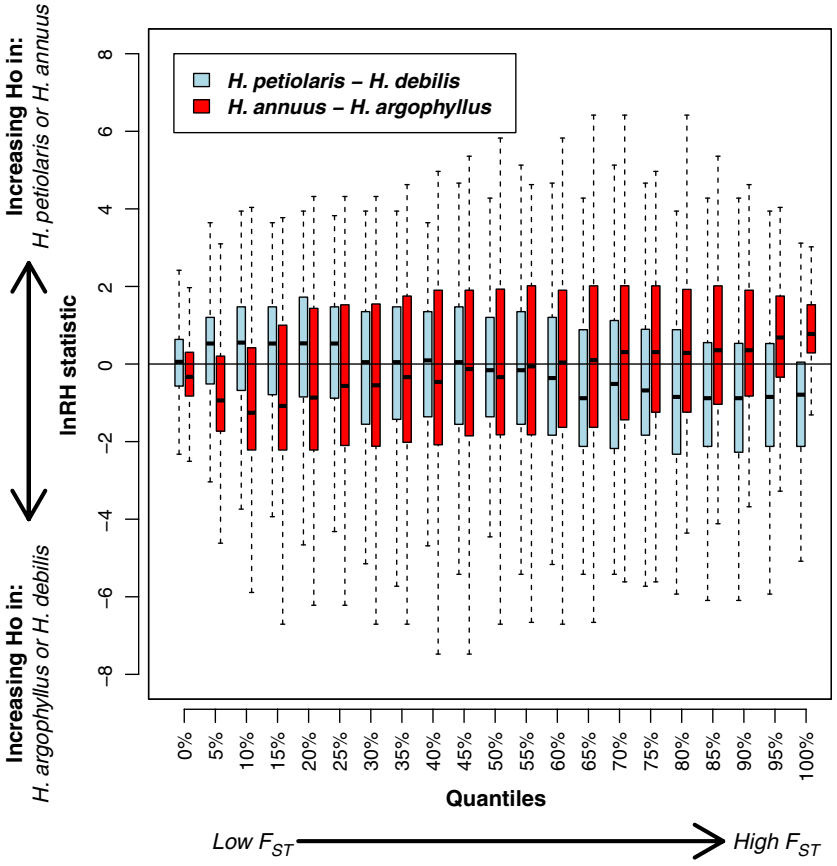

Fig. 5 lnRH statistic for each quantile. Genes were ranked according to F_{ST} for each species pair and split into twenty groups (quantiles), each representing 5% of the ranked data. The most divergent genes showed contrasting patterns of lnRH in the *Helianthus petiolaris – H. debilis* and *H. annuus – H. argophyllus* species pairs.

Table 3 Overrepresentation of Gene Ontology categories for shared genes showing with either high (top 15%) or low (bottom 15%) levels of divergence (corrected *P*-value < 0.05) in the *Helianthus annuus – H. argophyllus* versus *H. petiolaris – H. debilis* comparison, calculated using ErmineJ (Lee *et al*. 2005). *Number of probes* column corresponds to the number of *Helianthus* genes in the GO category. *Number of genes* column corresponds to the unique number of annotated *A. thaliana* genes these probes correspond to.

	Helianthus annuus – H. argophyllus versus *H. petiolaris – H. debilis* comparison				
Name	ID	Number of Probes	Number of Genes	*P*-value	Corrected *P*-value
High divergence genes					
Regulation of long-day photoperiodism, flowering	GO:0048586	17	7	1.14e-04	0.046
Regulation of photoperiodism, flowering	GO:2000028	22	8	1.80e-04	0.036
Low divergence genes					
Maintenance of inflorescence meristem identity	GO:0010077	15	6	7.73e-05	0.03

species pairs often act on the same genes, but frequently via different independent mutations.

To explore the nature of the selective forces leading to these patterns, we quantified patterns of repeatability for different classes of genes, going from the most divergent, to the least divergent ones. We observed that patterns of repeatability were not evenly distributed. Genes showing the highest (top 5%) or lowest (bottom 5%) divergence were much more often shared among species pairs than genes with less extreme levels of divergence, especially in the *H. petiolaris – H. debilis* versus *H. annuus – H. argophyllus* comparison (Fig. 4, see also Fig S1, Supporting information). We interpret these patterns as shared heterogeneity among diverging lineages in the strength and/or nature of selection faced by different genes. It is well established that all genes

are not equal in the eyes of evolution, with some genes repeatedly involved in local adaptation (Stern & Orgogozo 2009). This may, in turn, lead to the highly divergent genes being often shared among species pairs, especially when these are diverging along the same latitudinal gradient (Fig. 4). On the other hand, evolutionary conserved genes, such as genes coding for basic cellular processes under strong purifying selection, are more likely to be shared in closely related species than a random subset of genes, thus providing a plausible explanation for the quadratic relationships in Fig. 4. However, it is also possible that similar balancing selection pressure explains the left end part of the quadratic relationship, especially as low divergence shared genes seem to be missing in the two nonparallel comparisons in Fig. 4. If so, this would imply that balancing selection is acting on similar genes in the parallel case (*H. petiolaris – H. debilis* versus *H. annuus – H. argophyllus* comparison), but not the other two. Finally, the near lack of correlations for genes showing intermediate levels of divergence can be explained by variation in the strength or direction of selection among taxa.

At the same time, it is noteworthy that symmetry in the strength of divergent selection was not maintained from north to south in the *H. annuus – H. argophyllus* and *H. petiolaris – H. debilis* comparisons. Contrasting lnRH statistics (Fig. 5) imply that in the *H. annuus – H. argophyllus* species pair, the high F_{ST} genes exhibit decreased polymorphism in the southern member of the pair, *H. argophyllus,* while in the *H. petiolaris – H. debilis* species pair, reduced polymorphism is found in the northern taxon, *H. petiolaris*. While this may appear paradoxical, different historical distributions and complex demographic processes are probably responsible for this seemingly idiosyncratic result. For example, introgression between *H. annuus* and *H. debilis* may have increased genetic variance for genes involved in local adaptation in these two species, compared with *H. argophyllus* and *H. petiolaris*. Indeed, *H. annuus* has extended its range during its recent evolutionary history into *H. debilis* habitat, where some of our samples originated. This range extension appears to be a result of adaptive introgression with *H. debilis* (Heiser 1951; Scascitelli *et al.* 2010) and presumably would lead to higher heterozygosity for a biased fraction of both the *H. debilis* and *H. annuus* genomes.

While members of each of the sister species pairs analysed here are allopatric, this does not exclude genes flow between species from different pairings. *Helianthus annuus* and *H. petiolaris* are largely sympatric and appear to have exchanged a large fraction of their genome for much of their recent evolutionary history (Kane *et al.* 2009). This may explain their relatively low level of genetic divergence (Renaut *et al.* 2013a) and in turn contribute to repeatable patterns of divergence in the *H. petiolaris – H. debilis* vs *H. annuus – H. argophyllus* comparison. However, this does not invalidate our conclusions regarding the role of divergent selection in shaping patterns of divergence. Instead, it implies that adaptive mutations can have an extra-specific origin (adaptive introgression) and also that gene flow may be limited to certain regions of the genome, such as, for example, collinear regions with high recombination rates.

Effects of genomic landscape on patterns of repeatability

Importantly, the highest level of repeatability was identified when we compared large genomic blocks, which can be comprised of several hundred SNPs. Therefore, the particular genomic location of a mutation or gene appears to be as important in predicting overall patterns of divergence as the function of the gene itself. This in turn would imply that genome architecture imposes strong evolutionary constraints on rates of adaptation, independent of the role of similar selection pressures (Lynch 2007). Strengthening the argument that genetic divergence can be partly explained by shared features of the local genomic landscape is the fact that repeatable patterns were identified in all three comparisons, despite the fact that selection pressures are likely quite different.

In contrast to the well-recognized role of natural selection in evolution, less attention is typically paid to other evolutionary forces, which can nevertheless have strong implications on the outcomes of evolution (Lynch 2007; Losos 2010). As Lynch (2007) points out, recombination and mutation rates are the biological properties most likely to influence evolvability. These interact with the specific population genetic environment, of which effective population size (N_e) is a determinant factor, to affect the evolutionary paths taken by populations by limiting the number of available mutations and by restricting the effects of natural selection (Stern & Orgogozo 2009). Below we discuss how these factors, while largely stochastic in the short term, can contribute to the patterns of repeatable genomic divergence observed, in addition to the role of natural selection.

First, we previously showed that recombination rate was highly heterogeneous along the sunflower genome and that regions of low recombination were associated with higher overall divergence (Renaut *et al.* 2013a). Reductions in recombination rates promote divergence by extending the effects of directional selection, which reduces diversity at linked neutral sites (Nachman & Payseur 2012). If heterogeneity in recombination rate is shared among the three species pairs analysed here, this

will invariably lead to similar, and by extension, repeatable patterns of divergence. At the present time, we only have a direct estimate of the recombination rate for *H. annuus*, given that it requires the integration of species-specific information from both a physical and a genetic map. While speculative, we strongly suspect that based on the recent divergence of the sunflowers species studied here, and previously published comparative genetic mapping data (Lai *et al.* 2005), genomic architecture and recombination rates should be broadly similar in all annual sunflowers and thus bias genetic divergence along a similar evolutionary path. On the other hand, large-scale chromosomal rearrangements that suppress interpopulational recombination and facilitate adaptive differentiation within these regions (Rieseberg 2001; Faria & Navarro 2010) may help explain why correlation coefficients differ among chromosomes (Fig. 3d). Some of these large-scale rearrangements may be shared among the species and thus could contribute to repeatable patterns of divergence in some instances. Unfortunately, at this point, we do not have high-resolution genetic maps for all the species analysed here. Current efforts are underway to address this question.

Second, neutral theory predicts that in the absence of gene flow, increases in mutation rate will lead to increases in sequence divergence through genetic drift (Kimura 1968). Accordingly, shared biases in the production of variation can lead unrelated species to produce the same patterns of genetic divergence (Losos 2011). As the three species pairs compared here are thought to have evolved in conditions of reduced gene flow due to geographic isolation and restricted contact zones (Rogers *et al.* 1982), variability in mutation rate is expected to influence divergence. In fact, in all three species pairs, the rate of synonymous divergence (d_s), a proxy for the neutral mutation rate (Baer *et al.* 2007), was positively correlated with genetic divergence (F_{ST}) along the genome (Fig. S4, Supporting information). More importantly, d_s also displayed the same level of correlation among species pairs (Fig. S5, Supporting information). Therefore, in a similar fashion to the effect of recombination rate discussed above, increased mutation rate leads to increased divergence, and given that heterogeneity in mutation rate is shared among lineages, to repeatable patterns of divergence.

Conclusion

The genetic basis of recent evolution in sunflowers thus appears to be predictable (Stern & Orgogozo 2009). At the same time, we must not forget that many aspects of biology that may, at first sight, appear to be the result of natural selection may also be severely constrained by nonadaptive processes. Genomic architecture, gene structure and evolvability are difficult to make sense of without invoking the nonadaptive forces of drift and mutation and therefore may largely be indirect by-products of processes operating at higher levels of genome organization (Lynch 2007). Especially in this era of next generation sequencing, it is imperative to move beyond approaches that limit exploration to a priori expectations based on previously suspected candidate genes. One such approach is to first identify, in an unbiased manner, loci involved in adaptation and then ask whether the proportion of changes that are shared is greater than a neutral expectation. This will also require more comprehensive comparative genome scans, which will help to disentangle the different factors modulating genomic evolution.

Acknowledgements

We wish to thank Axios Review (axiosreview.org) for providing presubmission reviewing of an earlier version of this manuscript. This work was supported by an NSERC Postdoctoral Fellowship to SR, an NSERC Grant (327475) to LHR and a NSERC Doctoral Fellowship to GLO.

References

Andrew RL, Rieseberg LH (2013) Divergence is focused on few genomic regions early in speciation: incipient speciation of sunflower ecotypes. *Evolution*, **67**, 2468–2482.

Arendt J, Reznick D (2008) Convergence and parallelism reconsidered: what have we learned about the genetics of adaptation? *Trends in Ecology & Evolution*, **23**, 26–32.

Baer CF, Miyamoto MM, Denver DR (2007) Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nature Reviews Genetics*, **8**, 619–631.

Bernatchez L, Renaut S, Whiteley AR *et al.* (2010) On the origin of species: insights from the ecological genomics of lake whitefish. *Philosophical Transactions Of The Royal Society B-Biological Sciences*, **365**, 1783–1800.

Blackman BK, Michaels SD, Rieseberg LH (2011) Connecting the sun to flowering in sunflower adaptation. *Molecular Ecology*, **20**, 3503–3512.

Bowers JE, Bachlava E, Brunick RL *et al.* (2012) Development of a 10,000 locus genetic map of the sunflower genome based on multiple crosses. *G3: Genes | Genomes | Genetics*, **2**, 721–729.

Brady KU, Kruckeberg AR, Bradshaw HD Jr (2005) Evolutionary ecology of plant adaptation to serpentine soils. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 243–266.

Conte GL, Arnegard ME, Peichel CL, Schluter D (2012) The probability of genetic parallelism and convergence in natural populations. *Proceedings Of The Royal Society Of London Series B-Biological Sciences*, **279**, 5039–5047.

Darwin C (1859) *Darwin: The Origin Ofspecies - Google Scholar*. Murray, London.

Elmer KR, Meyer A (2011) Adaptation in the age of ecological genomics: insights from parallelism and convergence. *Trends in Ecology & Evolution*, **26**, 298–306.

Eyre-Walker A, Keightley PD (1999) High genomic deleterious mutation rates in hominids. *Nature*, **397**, 344–347.

Faria R, Navarro A (2010) Chromosomal speciation revisited: rearranging theory with pieces of evidence. *Trends in Ecology & Evolution*, **25**, 660–669.

Goudet J (2005) HIERFSTAT, a package for R to compute and test hierarchical F-statistics. *Molecular Ecology Notes*, **5**, 184–186.

Grabherr MG, Haas BJ, Yassour M *et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, **29**, 644–652.

Heiser CB Jr (1951) Hybridization in the annual sunflowers: *Helianthus annuus* × *H. debilis* var. cucumerifolius. *Evolution*, **5**, 42–51.

Hoekstra HE (2006) Genetics, development and evolution of adaptive pigmentation in vertebrates. *Heredity*, **97**, 222–234.

Hoekstra HE, Hirschmann RJ, Bundey RA, Insel PA, Crossland JP (2006) A single amino acid mutation contributes to adaptive beach mouse color pattern. *Science*, **313**, 101–104.

Huson DH (2005) Application of phylogenetic networks in evolutionary studies. *Molecular Biology And Evolution*, **23**, 254–267.

Jones FC, Grabherr MG, Chan YF *et al.* (2012) The genomic basis of adaptive evolution in threespine sticklebacks. *Nature*, **484**, 55–61.

Kane NC, King MG, Barker MS *et al.* (2009) Comparative genomic and population genetic analyses indicate highly porous genomes and high levels of gene flow between divergent *Helianthus* species. *Evolution*, **63**, 2061–2075.

Kane N, Gill N, King M *et al.* (2011) Progress towards a reference genome for sunflower. *Botany*, **89**, 429–437.

Kawakami T, Morgan TJ, Nippert JB *et al.* (2011) Natural selection drives clinal life history patterns in the perennial sunflower species, *Helianthus maximiliani*. *Molecular Ecology*, **20**, 2318–2328.

Kimura M (1968) Evolutionary rate at the molecular level. *Nature*, **217**, 624–626.

Lai Z, Livingstone K, Zou Y *et al.* (2005) Identification and mapping of SNPs and candidate genes in sunflower: towards a functional map. *Theoretical and Applied Genetics*, **111**, 1532–1544.

Lai Z, Zou Y, Kane NC *et al.* (2012) Preparation of normalized cdna libraries for 454 titanium transcriptome sequencing. In: *Data Production and Analysis in Population Genomics*, pp. 119–133. Humana Press. New York, New York.

Lee HK, Braynen W, Keshav K, Pavlidis P (2005) BMC Bioinformatics | Full text | ERMINEJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, **6**, 269.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li H, Handsaker B, Wysoker A *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Liu A, Burke JM (2006) Patterns of nucleotide diversity in wild and cultivated sunflower. *Genetics*, **173**, 321–330.

Losos JB (2010) Adaptive radiation, ecological opportunity, and evolutionary determinism. *The American Naturalist*, **175**, 623–639.

Losos JB (2011) Convergence, adaptation, and constraint. *Evolution*, **65**, 1827–1840.

Lynch MM (2007) The frailty of adaptive hypotheses for the origins of organismal complexity. *PNAS*, **104** (Suppl 1), 8597–8604.

Manceau M, Domingues VS, Linnen CR, Rosenblum EB, Hoekstra HE (2010) Convergence in pigmentation at multiple levels: mutations, genes and function. *Philosophical Transactions Of The Royal Society B-Biological Sciences*, **365**, 2439–2450.

Nachman MW, Payseur BA (2012) Recombination rate variation and speciation: theoretical predictions and empirical results from rabbits and mice. *Philosophical Transactions Of The Royal Society B-Biological Sciences*, **367**, 409–421.

Nadeau NJ, Jiggins CD (2010) A golden age for evolutionary genetics? Genomic studies of adaptation in natural populations. *Trends In Genetics*, **26**, 484–492.

Olsson K, Agren J (2002) Latitudinal population differentiation in phenology, life history and flower morphology in the perennial herb *Lythrum salicaria*. *Journal Of Evolutionary Biology*, **15**, 983–996.

R Core Team (2012) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ralph P, Coop G (2010) Parallel adaptation: one or many waves of advance of an advantageous allele? *Genetics*, **186**, 647–668.

Renaut S, Grassa C, Moyers B, Kane N, Rieseberg L (2012) The population genomics of sunflowers and genomic determinants of protein evolution revealed by RNAseq. *Biology*, **1**, 575–596.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013a) Genomic islands of divergence are not affected by geography of speciation in sunflowers. *Nature Communications*, **4**, 1827.

Renaut S, Grassa CJ, Yeaman S *et al.* (2013b) Data from: Genomic islands of divergence are not affected by geography of speciation in sunflowers. Dryad Digital Repository. Available from: http://dx.doi.org/doi:10.5061/dryad.9q1n4.

Rieseberg LH (2001) Chromosomal rearrangements and speciation. *Trends in Ecology & Evolution*, **16**, 351–358.

Rieseberg LH, Raymond O, Rosenthal DM *et al.* (2003) Major ecological transitions in wild sunflowers facilitated by hybridization. *Science*, **301**, 1211–1216.

Roesti M, Hendry AP, Salzburger W, Berner D (2012) Genome divergence during evolutionary diversification as revealed in replicate lake–stream stickleback population pairs. *Molecular Ecology*, **21**, 2852–2862.

Rogers C, Thompson T, Seiler GJ (1982) *Sunflower Species of the United States*. National Sunflower Association, Bismarck, North Dakota.

Römpler H, Rohland N, Lalueza-Fox C, Willerslev E (2006) Nuclear gene indicates coat-color polymorphism in mammoths. *Science*, **313**, 62.

Rosenblum EB, Hoekstra HE, Nachman MW (2004) Adaptive reptile color variation and the evolution of the MC1R gene. *Evolution*, **58**, 1794–1808.

Scascitelli M, Whitney KD, Randell RA *et al.* (2010) Genome scan of hybridizing sunflowers from Texas (Helianthus annuus and H. debilis) reveals asymmetric patterns of introgression and small islands of genomic differentiation. *Molecular Ecology*, **19**, 521–541.

Schlotterer C, Dieringer D (2005) A novel test statistic for the identification of local selective sweeps based on microsatellite gene diversity. In: *Selective Sweep* (ed. Nurmin-sky D), pp. 55–64. Landes Bioscience, Georgetown, Texas.

Smith SD, Rausher MD (2011) Gene loss and parallel evolution contribute to species difference in flower color. *Molecular Biology And Evolution*, **28**, 2799–2810.

Stern DL, Orgogozo V (2009) Is genetic evolution predictable? *Science*, **323**, 746–751.

Strasburg JL, Kane NC, Raduski AR *et al.* (2011) Effective population size is positively correlated with levels of adaptive divergence among annual sunflowers. *Molecular Biology And Evolution*, **28**, 1569–1580.

Strasburg JL, Sherman NA, Wright KM *et al.* (2012) What can patterns of differentiation across plant genomes tell us about adaptation and speciation? *Philosophical Transactions Of The Royal Society B-Biological Sciences*, **367**, 364–373.

Streisfeld MA, Rausher MD (2009) Genetic changes contributing to the parallel evolution of red floral pigmentation among Ipomoeaspecies. *New Phytologist*, **183**, 751–763.

Theron E, Hawkins K, Bermingham E, Ricklefs RE, Mundy NI (2001) The molecular basis of an avian plumage polymorphism in the wild. *Current Biology*, **11**, 550–557.

Timme RE, Simpson BB, Linder CR (2007) High-resolution phylogeny for *Helianthus* (Asteraceae) using the 18S-26S ribosomal DNA external transcribed spacer. *American Journal of Botany*, **94**, 1837–1852.

Turner TL, Levine MT, Eckert ML, Begun DJ (2008) Genomic analysis of adaptive differentiation in *Drosophila melanogaster*. *Genetics*, **179**, 455–473.

Turner TL, Bourne EC, Von Wettberg EJ, Hu TT, Nuzhdin SV (2010) Population resequencing reveals local adaptation of Arabidopsis lyrata to serpentine soils. *Nature Genetics*, **42**, 260–263.

Weir B, Cockerham CC (1984) Estimating F-statistics for the analysis of population-structure. *Evolution*, **38**, 1358–1370.

Wood TE, Burke JM, Rieseberg LH (2005) Parallel genotypic adaptation: when evolution repeats itself. *Genetica*, **123**, 157–170.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology And Evolution*, **24**, 1586–1591.

S.R. and L.H.R. designed the study. S.R. analyzed the data. S.R. and G.L.O. gathered the data. S.R. wrote the manuscript, while G.L.O. and L.H.R. provided comments.

Data accessibility

The SNP tables, F_{ST} values for the SNPs, genes and genome as well as d_s values were deposited in the Dryad Digital Repository: doi:10.5061/dryad.465v3. R scripts and readme files used to generate outputs are accessible on github (https://github.com/seb951/helianthus_repeatability_genomic_divergence). Raw data analysed in this study are accessible via the Sequence Read Archive at NCBI under the Accession codes SRX264548 to SRX264569, SRX264812 to SRX264817, SRX264824 to SRX264825, SRX264836 to SRX264912.

Supporting information

Additional supporting information may be found in the online version of this article.

Fig. S1 Heat map showing quantile (each representing 5% of the *F*ST distribution) ranks for all three comparisons.

Fig. S2 Relationship between the number of observed versus expected shared markers and *F*ST.

Fig. S3 Correlation coefficients for each of the three species pairs of *F*ST against recombination rate calculated for *H annuus*.

Fig. S4 Correlation coefficient per species pairs of the rate of synonymous mutations per synonymous site [log(*d*s)] against *F*ST.

Fig. S5 Correlation coefficient among species pairs for the rate of synonymous mutation per synonymous site [log(*d*s)].

Table S1 Summary statistics (M : million).