Maintaining their genetic distance; limited gene flow between widely hybridising species of *Geum* with contrasting mating systems

Crispin Y. Jordan\textsuperscript{1,3,4}, Konrad Lohse\textsuperscript{1}, Frances Turner\textsuperscript{2}, Marian Thomson\textsuperscript{2}, Karim Gharbi\textsuperscript{3}, Richard A. Ennos\textsuperscript{3}

\textsuperscript{1}Institute of Evolutionary Biology, University of Edinburgh, Ashworth Laboratories, Charlotte Auerbach Rd, Edinburgh EH9 3FL, Scotland

\textsuperscript{2}Edinburgh Genomics, Ashworth Laboratories, Charlotte Auerbach Rd, Edinburgh EH9 3FL, Scotland

\textsuperscript{3}Present Address: University of Edinburgh, Hugh Robson Building, George Square, Edinburgh, EH8 9AG, Scotland

\textsuperscript{3}Corresponding author. E-mail: crispinjordan@gmail.com; Fax: 011 44 131 650 2872

Key words: Hybridization, introgression, coalescent, *Geum*, natural selection

Running title: Introgression and mating systems
Abstract

Mating system transition from outcrossing to selfing frequently gives rise to sister lineages with contrasting outcrossing rates. The evolutionary fate of such lineages depends on the extent to which they exchange genes. We measured gene flow between outcrossing *Geum rivale* and selfing *G. urbanum*, two sister species derived by mating system transition, which frequently hybridise. A draft genome was generated for *G. urbanum* and used to develop dd-RAD data scorable in both species. Coalescent analysis of RAD data from allopatric populations indicated that the two species diverged 2-3 Mya, and that long term gene flow between them has been very low (M=0.04). *G. rivale* showed greater genetic diversity in sympathy than allopatry, but genetic divergence between species was no lower in sympathy than allopatry, providing little evidence for recent introgression. Clustering of genotypes revealed that, apart from four early generation hybrids, individuals in sympatic populations fell into two genetically distinct groups with <1% admixture that corresponded exactly to their morphological species classification. Although our data suggest limited gene flow, we observed joint segregation of two putatively introgressed SNPs in *G. urbanum* populations that was associated with significant morphological variation; this provides tentative evidence for rare introduction of novel genetic diversity by interspecific gene flow. Our results indicate that despite frequent hybridisation, genetic exchange between *G. rivale* and *G. urbanum* has been very limited throughout their evolutionary history.
Introduction

A key factor influencing the evolutionary trajectory of lineages is their level of gene exchange with related taxa (Abbott et al. 2016). A reduction in gene flow facilitates speciation (Coyne & Orr 2004). Limited gene flow from a related taxon may increase the genetic variability of a species and enable acquisition of novel traits that enhance its evolutionary potential, while maintaining its distinctness (Paoletti et al. 2006). However, if gene flow is more extensive, species may fuse (Carney et al. 2000). Thus, quantifying gene exchange among lineages is crucial for understanding the birth, continued evolution and possible extinction of species.

Quantifying gene flow between lineages is particularly valuable across lineage pairs that share an evolutionary transition, allowing patterns of genetic exchange associated with that evolutionary transition to be recognised. In many plant genera speciation is associated with a mating system transition from outcrossing to selfing (Stebbins 1957; Igic et al. 2006; Wright et al. 2013; Barrett et al. 2014). Although sister species with contrasting mating systems are typically ecologically distinct, hybridisation often occurs between them where habitat isolation breaks down or the ranges of the two taxa meet (e.g. *Mimulus* (Vickery, 1978; Brandvain et al. 2014; Kenny & Sweigart 2016), *Rhinanthus* (Ducarme et al. 2010), *Centaurium* (Brys et al. 2013)).

The transition from outcrossing to selfing affects genome evolution per se (Wright et al. 2008) and is commonly accompanied by changes in floral display (Sicard & Lenhard 2011). Thus the transition can potentially lead both to post-zygotic barriers to interspecific gene flow, due to interactions between differently evolved outcrossed and selfed genomes, and to pre-zygotic barriers, through its effect on mating opportunities (Fishman & Wyatt, 1999; Hu, 2015). Quantitative measurement of introgression between sister selfing and outcrossing taxa is needed to determine the strength of these effects, and to ascertain whether they prevent integration of novel genetic material into either species. Introgression would be particularly important for highly inbreeding taxa whose adaptive
potential may be impaired by low genetic variability, and a reduced efficacy of positive and
background selection (Charlesworth et al. 1993; Glémin & Ronfort 2013; Arunkumar et al. 2015).

Here we quantify gene flow between *Geum rivale* and *G. urbanum*, sister species that differ in mating
system and associated floral display (Fig. 1). *Geum rivale*, which typically occupies open, moist
habitats, is predominantly outcrossed (outcrossing rate $t = 0.8$) and possesses a pendulous flower
adapted to bee pollination. In contrast, *G. urbanum*, a plant of shaded, well drained sites, is
predominantly self-fertilised ($t = 0.15$) and bears erect flowers adapted to fly pollination (Taylor
1997a, b; Ruhsam et al. 2010).

Allopatric populations of both species occur: *G. rivale* in tall herb montane communities in the UK,
and at high latitude sites elsewhere in Europe; *G. urbanum* in the extreme south east of the UK,
where rainfall is low, and at the southern extremes of its distribution in continental Europe (Taylor,
1997a, b). However, over most of their distributions in the UK and continental Europe the taxa are
sympatric. Here they can be found as single species populations in “pure” habitats or together in
“mixed” habitats. Typical mixed habitats include naturally wet woodlands and artificially disturbed
river banks (Waldren et al. 1989).

In mixed habitats the highly fertile F1 hybrid *G. x intermedium* is widely reported, associated with
hybrid swarms (Taylor 1997a). Detailed analysis of a single hybrid swarm indicated the presence of
both F1 hybrids and early generation backcrosses to *G. rivale* that produce fertile seeds by both
outcrossing and selfing (Ruhsam et al. 2011, 2013). The resulting recombinant offspring express no
detectable fitness reduction when grown in a benign environment, but some recombinant types are
not apparently present in the population of established adults in the hybrid swarm (Ruhsam 2013),
perhaps due to selection (Ruhsam et al. 2013). Thus, the mating events occurring in mixed habitats
are common and favourable for introgression, particularly via backcrossing to *G. rivale*. However, so
far there has been no assessment of whether this has led to significant gene exchange between the two species.

To quantify gene exchange between the two Geum taxa we adopt a comparative genomic approach. We first generate a draft genome of the inbreeding species, G. urbanum, which we use to develop a set of SNP markers common to the two species. These SNPs are then scored in allopatric populations of both species likely to have been geographically isolated for the past 5ky. We analyse these data in a coalescent framework to estimate the age of the two taxa and measure interspecific gene flow from the time at which the species formed to the time when allopatric populations became isolated. We also use the allopatric samples to identify species-specific diagnostic SNPs.

To measure recent gene flow between the Geum taxa, in the period since the isolation of the allopatric populations, we analyse an additional set of samples taken from a broad area of sympatry. We look for signals of recent genome-wide introgression by comparing nucleotide diversity and genetic differentiation between the allopatric and sympatric samples. If introgression has occurred, significant increase in genetic diversity within and reduction in genetic differentiation between species are expected in sympatric compared to allopatric samples. We applied a clustering approach to the SNP data to determine the extent of admixture of G. rivale and G. urbanum genomes within individuals. Finally, we score the sympatric sample for previously identified species-specific diagnostic SNPs to quantify hybridisation and introgression at the level of individual genotypes. We relate these estimates of the hybridity of individuals, based on molecular data, to the morphology of these individuals measured in a common environment.
MATERIALS AND METHODS

i. Sampling of Allopatric Populations

Four *G. rivale* individuals were sampled from each of three high elevation tall herb communities located on three distinct Scottish mountain ranges above the altitudinal limits of *G. urbanum* (Table S1; Taylor 1997b). One individual was also sampled from a single population in each of northern Sweden and Iceland where *G. urbanum* is absent (Table S1; http://linnaeus.nrm.se/flora).

Two *G. urbanum* plants were sampled from each of 10 populations in south-east England (Table S1), a region where *G. rivale* is absent due to lack of suitable habitats (Preston et al. 2002). Single individuals of *G. urbanum* were also sampled from two areas in Europe where *G. rivale* does not occur, namely Portugal and south-west France (Table S1; http://linnaeus.nrm.se/flora). 0.5 grms young leaf tissue was removed from sampled plants and stored dry in silica gel prior to DNA extraction.

ii. Sampling of Sympatric Populations

The Botanical Society of Britain and Ireland (BSBI) database was searched for the location of sites containing *G. rivale* alone (15), *G. urbanum* alone (14), or both taxa (10) in eight contiguous 10km x 10km squares within vice county 81 in the British Isles (Berwickshire) (Fig. 1, Table S2) (http://bsbidb.org.uk/maps/). At each site a single cutting was collected in August 2013 from each of two plants more than 5m apart, for each of the *Geum* taxa present. Field identification employed floral, fruit and vegetative characters. Newly emerging leaves (0.1 g.) were taken from each plant and frozen (-80°C) prior to DNA extraction. Cuttings were rooted in 20cm diameter pots containing
commercial potting compost. Ninety six of the 98 cuttings were rooted successfully. Plants were overwintered outside in a randomised array in Edinburgh to induce flowering in spring 2014.

**Development of SNP markers via ddRAD**

**i. DNA extraction**

DNA was extracted from dried samples using a modification of Doyle & Doyle (1990)'s CTAB protocol, while fresh and frozen material was extracted using the DNeasy plant mini prep kit (Qiagen) following the manufacturer’s protocol.

**ii. Development of draft Genome**

We developed a draft genome for the inbreeder, *G. urbanum* to facilitate SNP genotyping with double-digest Restriction Associated DNA sequencing markers (ddRAD; Peterson et al. 2012). We identified the individual within our Punnetts Town sample (Table S1) that was most homozygous at a set of 10 microsatellite loci (Arens et al. 2004). Using this sample we extracted 17μg of DNA from 150mg of young leaves. Details of the genome sequencing and assembly, which was conducted by Edinburgh Genomics, are given in the Supplementary Materials.

To identify SNPs within the draft genome and estimate coverage per scaffold, trimmed reads were mapped back to assembled scaffolds using bwa mem version 0.7.5a with -M option (Li & Durbin 2009). 94.0% of reads were aligned to scaffolds. Samtools (Li et al. 2009) version 0.1.18 was used to call variants.

**iii. ddRAD library prep and sequencing**

DNA quality was evaluated using the E-Gel Precast Agarose Electrophoresis System (ThermoFisher Scientific), and samples quantified using the Qubit dsDNA BR Assay Kit, (ThermoFisher Scientific).
Sample preparation, library construction and PCR amplification for double digest RAD sequencing was modified from Peterson et al. (2012) (for full details see Supplementary Materials). Superpools of PCR products were sequenced either on an Illumina HiSeq 2500 using 125 based paired-end reads in high output mode (v4 chemistry) or an Illumina MiSeq using 150 base paired-end reads (v2 chemistry).

iv. Quality filtering and preparation of reads from ddRAD

The bioinformatics and analysis pipeline for all ddRAD data is summarized in Figure S1. To match the read lengths produced by MiSeq and HiSeq Illumina technologies for ddRAD analyses, we used fastx trimmer (http://hannonlab.cshl.edu/fastx_toolkit/) to trim MiSeq reads to 125bp (STACKS requires reads of equal length). We then de-multiplexed and filtered reads for quality using process_radtags (STACKS v 1.21; Catchen et al. 2011, 2013). This process removed reads with an uncalled base, and those with an average quality score below 20 over a sliding window comprising 15% of a read. Finally, we again trimmed reverse reads to 117bp with fastX trimmer to produce forward and reverse reads of equivalent length after the 8bp barcode was removed from the forward reads.

v. Aligning reads to draft genome

Both Geum species are ancient hexaploids (2N = 42), with disomic inheritance (Smedmark et al. 2003; Vandepitte et al. 2007; Ruhsam et al. 2010). Mapping ddRAD sequences from the two Geum species must account for divergence between species to allow reads from both species to map to the G. urbanum draft genome. However, allowing an excessive number of mismatches can cause reads to map to multiple loci in ancient polyploid taxa, like Geum. We explored the effect of varying the number of allowed mismatches (M) on alignments for one allopatric individual from each species (Punnetts Town, G. urbanum; Ben Lui, G. rivale; both sequenced with Illumina HiSeq).
For a range of M values, we used GSnap (version 2014-12-16; Wu & Watanabe 2005) to map reads to the draft genome. We required that reads map uniquely, specified an indel penalty of two, and did not allow terminal alignments. For the G. *urbanum* sample, between 64.86% to 67.74% of paired-end (PE) ddRAD reads mapped as proper pairs when we allowed one to nine mismatches (M=1 to 9); however the percentage of reads that aligned jumped to 82.84% for M=10. Similarly, the percentage of G. *rivale* PE reads that aligned in proper pairs increased gradually from 39.21% to 58.16% over M=2 to M=9, with a large increase to 79.93% for M=10. Therefore, the structure of the *Geum* genome (e.g., due to polyploidy and/or repetitive sequences) appears to present different environments for aligning ddRAD data at a threshold genetic diversity of ~8.55% (i.e., 10/117 * 100, where 117 equals the read length). Consequently, we performed all subsequent alignments as above, setting M=6 to avoid this threshold but also allow many reads from *G. rivale* to map to the heterospecific draft genome.

vi. **Assembling radtags and calling SNPs**

We used STACKS' ref_map pipeline (v 1.21; Catchen *et al.* 2011, 2013; Hohenlohe *et al.* 2010) to assemble the aligned reads into radtags. Next, we used STACKS' rxstacks module to correct genotype calls based on population-level genotype information using a bounded (error rate) model with the parameters conf_lim and bound_high equal to 0.25 and 0.1, respectively (see http://catchenlab.life.illinois.edu/stacks/manual/).

vii. **Identifying and filtering for paralogs**

Ancient hexaploidy complicates analyses of *Geum* because reads from paralogous genome regions could map to identical sites. As a strict filter for identifying genome regions that potentially attract paralogous reads, we used STACKS’ populations module to calculate observed heterozygosity and $F_{IS}$ at each identified SNP. Thereafter custom scripts were used to collate data and restrict further analysis to scaffolds free of SNPs that exhibited either excess heterozygosity (>0.5) or negative $F_{IS}$. 
The heterozygosity and $F_\text{ST}$ calculations for this purpose only considered UK allopatric *Geum* samples that comprised four individuals from each of three Scottish ‘allopatric’ *G. rivale* populations (12 *G. rivale* individuals), and one randomly chosen individual from each of the 10 Southern England *G. urbanum* populations (Table S1). Nucleotide-level heterozygosity and $F_\text{ST}$ were calculated separately for *G. rivale* and *G. urbanum*; we assumed each set of samples was from a single, large UK population, despite population structure being evident in both species (see Results). Population structure reduces the number of heterozygotes compared to a single panmictic population: therefore, our filters should preferentially remove genomic regions with excess heterozygosity due to paralogy.

**Population Genetic Analysis of Allopatric Populations**

*i. Patterns of Polymorphism within and between species*

We characterized polymorphism within and between the *Geum* species using the 12 allopatric UK *G. rivale* samples plus one from each of Iceland and Sweden, and 10 UK *G. urbanum* samples (one from each allopatric UK population) plus one each from France and Portugal (Table S1). We used STACKS’ populations module to identify SNPs that are polymorphic within at least one species or show a fixed difference between species. This analysis only considered radtags present in all 26 individuals described above.

*ii. Inbreeding coefficients and population differentiation*

We applied the STACKS’ populations module to estimate the inbreeding coefficients within each population ($F_\text{IS}$) and measure differentiation among populations ($F_\text{ST}$) for each species using data from the three allopatric *G. rivale* populations in Scotland (4 individuals/pop) and the ten allopatric *G. urbanum* populations.
urbanum populations in England (2 individuals/pop). Estimates of $F_{\text{IS}}$ and $F_{\text{ST}}$ only considered radtags present in both species, and that were present in all individuals of a given analysis.

iii. Identification of species-specific SNPs

We used alternately fixed SNPs in allopatric populations of the two taxa (identified in (i.) above) to provide an initial list of species-specific SNPs. To minimize linkage between diagnostic SNPs and obtain an estimate of introgression across the whole genome, we selected one (the first) alternately fixed SNP per scaffold for the introgression analysis. However, we note that by using the first SNP per scaffold, we bias our data against (larger) well-assembled scaffolds. Therefore, we repeated our analyses of introgression using all available species-diagnostics SNPs to test for an effect of such bias, and obtained qualitatively identical results (not shown).

iv. Coalescent analysis of gene flow during lineage divergence

We used an analytic likelihood framework to assess the support for alternative models of divergence between G. rivale and G. urbanum with and without gene flow. The method is described in Lohse et al. (2016). Briefly, the analysis is based on a single diploid individual for each species and considers the blockwise site frequency spectrum, i.e. the joint frequencies of four polymorphism types (as in “Patterns of Polymorphism within and between species”, above) in short blocks of sequence: i) heterozygous sites exclusive to G. rivale, ii) heterozygous sites exclusive to G. urbanum, iii) heterozygous sites shared by both species and iv) fixed differences between species. We counted these site types within 117bp radtags (block), and treated each radtag as an independent block. For randomly mating populations, the polymorphisms at each block represent an independent outcome of the coalescent process, which is a function of the species’ history. Since STACKs ignores RAD tags that are monomorphic, we conditioned the likelihood on only observing variable blocks. To do this, we normalized the probability of each blockwise mutational configuration by $1 - p_{\text{IBS}}$, where $p_{\text{IBS}}$ is the probability of identity in state for a block.
We primarily wish to test for historical introgression between *G. rivale* and *G. urbanum*. Therefore, we compared three models: (1; “div<sub>2</sub>”) species diverged at time *T* with no introgression, (2; “IM<sub>_u</sub>_r”) a history of divergence with gene flow from *G. urbanum* to *G. rivale* and (3; “IM<sub>_r</sub>_u”) divergence with gene flow from *G. rivale* to *G. urbanum* (4 parameters). All models assume instantaneous species divergence and a constant *N<sub>e</sub>* within taxa. We also constrained the effective size of the ancestor to equal that of *G. rivale*. To capture the decrease of *N<sub>e</sub>* expected to result from the transition to inbreeding (Charlesworth & Wright 2001), we allowed *N<sub>e</sub>* to differ between *G. urbanum* and the other taxa (the ancestor of *G. rivale* and *G. urbanum* is most likely outcrossing given the rarity of transitions from self-compatibility to self-incompatibility; Igic et al. 2006). Migration was modelled as a constant rate *M* = 4*N<sub>anc</sub> x m individuals per generation. We converted estimates of *T* into years using *t* = *T* x 2*N* x g where *g* is generation time and *N* = θ/4μ. We assumed a mutation rate estimated for *Arabidopsis thaliana*, μ = 7 x 10<sup>-9</sup> (Ossowski et al. 2010), and an average generation time across the two species of 3 years (Taylor 1997a, b) (generations are likely longer in *G. rivale* than in *G. urbanum*).

Inbreeding complicates this analysis because it reduces genetic diversity within individuals relative to the population level. To minimize the confounding effect of inbreeding on the analyses, we initially chose the most outbred individuals for analysis. We used a recently outcrossed *G. urbanum* individual (see Results; population Mill Wood), which had genetic diversity (*π* = 0.0009 ± 0.0001) similar to the UK allopatric *G. urbanum* populations as a whole (see Results) and so may adequately represent the diversity of this species in the UK. “Leaky” self-incompatibility also introduces variation in inbreeding among individuals of *G. rivale*. Therefore, we conducted separate analyses that paired the outbred *G. urbanum* Mill Wood individual with 3 allopatric UK samples of *G. rivale* that span a range of heterozygosity to examine how inbreeding in *G. rivale* affects our conclusions.
Maximum likelihood estimates under each model were obtained in *Mathematica* v. 10.2 (see File S2 of Lohse *et al.* 2016). To estimate 95% CI for M and T, we obtained discretized marginal support (logarithm of the likelihood) curves for these parameters (maximising the likelihood for all other parameters at each point).

Population Genetic Comparison of Allopatric and Sympatric Populations

**i. Genetic diversity in allopatric and sympatric populations**

To determine whether potential genetic exchange in sympatric populations resulted in an increase in genetic diversity, we estimated genetic diversity (\(\pi\)) in UK allopatric and Berwickshire samples (pooling Berwickshire’s ‘pure’ and ‘mixed’ populations) of each species using STACKS’ populations module (providing mean and standard error, S.E.). The analysis considered radtags present in at least half of individuals for each species/population type (allo- vs sympatric) combination; we obtained qualitatively similar results when we restricted the analysis to radtags present in all individuals (not shown). The analysis excluded four individuals in the Berwickshire sample identified as obvious early generation hybrids or backcrosses (see Results). This analysis therefore involved 12 allopatric *G. rivale*, 20 allopatric *G. urbanum*, 48 sympatric *G. rivale* and 45 sympatric *G. urbanum* samples.

**ii. Genetic differentiation between taxa in allopatry and sympathy**

To test whether potential gene exchange resulted in less differentiation between the species in sympatric populations, we calculated genetic differentiation \(d_{xy} = \Sigma[(PX \ast (1-PY)) + ((1-PX) \ast PY)] / n\) between the two species, separately for allopatric and sympatric samples: \(P\) represents the frequency of a focal allele in the sample of species X and Y, and \(n\) is the sequence length (Nei & Li 1979; Cruickshank & Hahn 2014). These calculations considered the same samples as used to calculate genetic diversity (\(\pi\)), described above. We estimated \(d_{xy}\) for each scaffold, which we
assume to be independent with respect to linkage and determined the mean $d_{xy}$ and its S.E. across scaffolds. All analyzed radtags were present in at least 12 individuals for each species/population type combination, and the same scaffolds were analyzed in allopatric and sympatric estimates of $d_{xy}$ (n=418 scaffolds).

**Genetic Analysis of Hybrids and Introgression in Sympatric Populations**

*i. Cluster analysis using fastSTRUCTURE*

As a first approach to analyzing introgression between the two *Geum* taxa in sympatry, we used genotypic clustering implemented in fastSTRUCTURE (Raj et al. 2014). Our analysis assumed that samples derived from two populations (i.e., $K=2$) corresponding to the two *Geum* species, with the possibility of genetic admixture of individuals. The SNP data was derived from a STACKS analysis of all 92 successfully genotyped individuals in the sympatric Berwickshire sample together with data from the British and European allopatric populations of the two species. We analyzed two datasets that differed in degree of filtering: (1) SNPs were present in any fraction of individuals, which led to greater numbers of SNPs but missing values for some individuals, and (2) all SNPs were present in all individuals. Both analyses used SNPs that had been filtered for paralogs, and considered only a single SNP per scaffold to minimize the effect of linkage between SNPs.

**ii. Identifying hybrids and introgressed individuals using species-specific SNPs**

In our second approach to analyzing recent introgression we used custom scripts and the species-specific SNPs identified in (iii.) (above) to estimate the fraction of alleles in each individual within the sympatric Berwickshire sample that is *G. rivale* in origin (Hybrid Index). SNPs with either un-callable genotypes (based on STACKS' likelihood algorithm) or third alleles (e.g., due to sequencing or alignment error) were excluded from the calculations. We simultaneously tabulated the frequency of SNPs that were heterozygous for the species-diagnostic alleles in each individual.
In general, our analyses did not specify minimum coverage because STACKS accounts for coverage when calling genotypes (or leaving a genotype uncalled) (Hohenlohe et al. 2010; Catchen et al. 2013). However, to check whether specifying minimum coverage would alter our results, we repeated our analyses, requiring a minimum coverage of 20 in all analysed individuals when running fastSTRUCTURE and creating the panel of species-diagnostic SNPs. This quartered the number of species-diagnostic SNPs, but yielded qualitatively identical results (not shown).

**Analysis of Morphological Variation in Sympatric Populations**

*i. Measurement of morphological variation in sympatric populations*

From the beginning of April 2014 plants grown from cuttings from sympatric populations were monitored weekly for flowering. From each plant a newly opened flower and the stipule located on the flowering stem immediately below the flower were sampled. The following characters, which discriminate between *G. rivale* and *G. urbanum*, were measured: angle at which flowers are held (degrees from vertical) (FA), petal length (PL) (mm), petal width (PW) (mm), petal shape (PS) (proportional height of widest part of petal)), sepal length (SL) (mm), stamen number (SN), stipule length (STL) (mm) and stipule width (STW) (mm) (Ruhsam et al. 2011, 2013).

*ii. Statistical analysis of morphological data*

Principal Component Analysis (PCA) was conducted on the total sympatric sample using data on all eight morphological characters measured. Morphologically distinct groupings were recognised on scatterplots of the first two principal component scores and related to parental species and hybrid classes defined genetically by species-specific SNPs. PCA based on the same characters was also used to summarise morphological variation separately within *G. urbanum*. GLMs were used to test for the effect of two SNP variants, putatively jointly introgressed from *G. rivale* (see results), on the first two principal component scores for *G. urbanum*. All statistical analyses were conducted in Minitab 16.
RESULTS

Development of SNP markers via ddRAD

i. Draft Genome

The genome assembly contained a total of 170,030 scaffolds, with an N50 of 24.6Kb and total assembly size of 1.2Gb. In order to identify signs that scaffolds represent multiple copies of the genome, the distributions of coverage per scaffold and percent variant bases per scaffold was assessed. Scaffolds of length less than 10Kb were excluded, leaving 32,182 scaffolds (with a total length of 909 Mb). The distribution of percent variant bases per scaffold is shown in Figure S2, and the distribution of coverage per scaffold is shown in Figure S3. Coverage per scaffold follows a roughly normal distribution, which would be expected if the scaffolds mostly represented the same number of copies of the genome.

Core genic regions were well assembled, and appeared to be present in approximately three (haploid) copies as expected for an ancient hexaploid. We searched our assembled genome for 248 ultra-conserved eukaryotic genes (CEGs), listed by Parra et al. (2007), using their Core Eukaryotic Genes Mapping Approach. We identified 93% and 97% of the core genes that were complete or partially complete, respectively. On average, complete and partially complete CEGs were represented by 3.39 and 3.82 orthologs per CEG, respectively, with at least 90% of CEGs represented by more than one ortholog.

ii. ddRAD tags

Following quality filtering, 2.7 * 10^7 reads remained in the MiSeq data, and between 6.0 * 10^7 and 7.3 * 10^7 reads remained that derived from the five HiSeq libraries. Following alignments, assembling radtags with STACKS, and applying corrections with rxstacks, our dataset included 230,356 radtags
for *G. rivale* and *G. urbanum*, collectively. However, coverage was highly stochastic. For example, only ~2% (4524) of radtags were represented in more than one half of our samples.

**iii. Identifying and filtering for paralogs**

When the STACKS’ populations module was used to analyse the raw SNP data from the allopatric populations of the inbreeding taxon *G. urbanum*, \(F_{IS}\) estimates were low and sometimes negative, which is unexpected for a highly inbreeding species. This suggests that paralogous reads have mapped to identical locations and thereby increased individual heterozygosity (see Table S4). We therefore applied our filter for paralogy (rej ecting 1344 scaffolds with SNPs that exhibit either excess heterozygosity (>0.5) or negative \(F_{IS}\)) and, unless specifically noted, we henceforth only present results from paralogy-filtered SNPs.

**Population Genetic Analysis of Allopatric Populations**

**i. Patterns of polymorphism within and between species**

The majority of SNPs in the dataset were polymorphic in *G. rivale* but invariant in *G. urbanum*. This type of polymorphic site occurred approximately four times more frequently than the reverse case (Table 1). 22% of SNPs were alternately fixed between the species (Table 1) and only 1.5% of SNPs were shared polymorphism (Table 1).

**ii. Inbreeding coefficients and population differentiation**

For the allopatric populations sampled in Britain we obtained \(F_{IS}\) estimates close to 0.25 for *G. rivale*, consistent with leaky self-incompatibility (Ruhsam et al. 2010), and \(F_{IS}\) greater than 0.9, consistent with very high selfing rates, for eight of the ten UK ‘pure’ *G. urbanum* populations (Table 2). Among the two *G. urbanum* populations with low \(F_{IS}\), the first population (Mill Wood \(F_{IS} = 0.0845\)), included one (of the two) sample(s) that was heterozygous at 92% of the 71 polymorphic SNPs analysed,
suggesting that this sample was derived from a recent outcrossing event between two diverged inbred lines. In the second population, Selwyn Wood, only one SNP was recorded as polymorphic, suggesting that this population may have been founded by few (possibly a single) highly selfing individuals.

Allopatric populations of *G. rivale* in Britain exhibited less population differentiation than *G. urbanum* populations (mean pairwise $F_{ST}$ 0.13 and 0.38 respectively), as expected from their contrasting breeding systems.

**iii. Coalescent analysis of gene flow during lineage divergence**

Table S5 summarizes the numbers of each polymorphism type and blocks analysed for the *G. urbanum* - *G. rivale* sample pairs. All three pairs had approximately 1400 SNPs distributed among ca. 660 blocks (i.e., radtags). Model comparisons for all three sample pairs reject a model of strict divergence, and suggest that introgression has occurred between the *Geum* species (Table 3), but at a very low rate (see below). For two sample pairs (involving *G. rivale* samples Ben Lui 1 and Ben Lawers 5), the model of gene flow from *G. rivale* to *G. urbanum* (i.e., IM$_{r\rightarrow u}$) fit the data significantly better than the model of strict divergence (i.e., div$_2$), whereas the model IM$_{u\rightarrow r}$ does not fit significantly better than div$_2$ (Table 3). Models that include gene flow (IM$_{r\rightarrow u}$ and IM$_{u\rightarrow r}$) also fit the data significantly better than the div$_2$ model for the third pair (involving sample Ben Lui 4), but IM$_{r\rightarrow u}$ and IM$_{u\rightarrow r}$ have effectively equal support (Table 3). Results from this latter pair likely differ from the former pairs because it includes a single block with a shared heterozygous site, while the other sample pairs lack shared heterozygous sites (Tables S5, S6). As Ben Lui 4 is likely the most heterozygous *G. rivale* sample (Table S5), we focus on this pair, but note that parameter estimates (Table S6) and general conclusions (i.e., support for very low introgression rates) are similar for all three sample pairs.
The three models (div$_2$, IM$_{ru}$, IM$_{ur}$) yield similar estimates of Ne and divergence time (Ben Lui 4 pair: Table 4; all three sample pairs: Table S6). In general, Ne of G. urbanum is a half to a quarter of that of G. rivale (and their common ancestor), and all three models suggest that the species diverged approximately 2 to 3 million years ago (Table 4; Table S6). Models IM$_{ru}$ and IM$_{ur}$ both suggest a low but significant long-term rate of effective gene flow (M ≈ 0.04), of approximately one migrant every 25 generations (Table 4; see Table S6 for M for additional sample pairs).

Population Genetic Comparison of Allopatric and Sympatric Populations

i. Genetic diversity in allopatric and sympatric populations

Estimates of genetic diversity (π) were based on 826 radtags, each 117 base pairs in length, totaling 96 642 base-pairs; as these radtags were the same as those used to develop species-diagnostic SNPs, they occur in both species. Genetic diversity was greater in G. rivale (mean π ± SE: ‘allopatric’ 0.0044 ± 0.0001; Berwickshire 0.0048 ± 0.0001) than G. urbanum (‘allopatric’ 0.0011 ± 0.0001; Berwickshire 0.0011 ± 0.0001), and was similar between population types for G. urbanum. However G. rivale genetic diversity was slightly greater for sympatric than allopatric samples. Since STACKS ignores monomorphic radtags, we note that these estimates of π should be interpreted in terms of relative, but not absolute diversity (see also Arnold et al. 2013).

ii. Genetic differentiation between taxa in allopatry and sympatry

Genetic differentiation between G. rivale and G. urbanum is similar for UK allopatric and sympatric samples (mean d$_{xy}$ (+/- SE) equals 0.0115 +/- 0.0005 and 0.0112 +/- 0.0005, respectively). Again, we note that, due to the fact that STACKS’ ignores monomorphic radtags, these estimates should be viewed as relative measures (and upper limits) of d$_{xy}$, and not as absolute estimates.
**Hybrids and Introgression in Sympatric Populations**

*i. Cluster analysis using fastSTRUCTURE*

fastSTRUCTURE analysis that considered SNPs present in either all (188 SNPs) or a proportion of (492 SNPs; results not shown) individuals analysed in the combined allopatric and sympatric populations yielded highly consistent results. Figure 2 illustrates fastSTRUCTURE results for analyses based on 188 SNPs. All except one of the 36 individuals in the allopatric populations, and all except 4 of the 96 individuals in the sympatric populations, show less than 1% admixture. In the sympatric population Q values (proportion of G. rivale genome) of the four individuals with substantial admixture are 0.496, 0.517, 0.663 and 0.935. On the basis of these Q values there is no evidence for significant introgression from G. rivale into G. urbanum. The substantially admixed individuals present are likely to be recently formed F1 or F2 hybrids, and different generations of backcrosses to G. rivale.

*ii. Identifying hybrids and introgressed individuals using species-specific SNPs*

To initially assess the ability of the species-specific SNPs to discriminate the species, we genotyped the 10 British allopatric G. urbanum samples that had not been used to generate species-diagnostic SNPs. 215 out of 220 diagnostic SNPs were homozygous for the G. urbanum allele in all individuals. Five SNPs were homozygous for the putative G. rivale allele in one or more of the 10 G. urbanum individuals, indicating that these SNPs were not fixed between species. They were removed from subsequent analysis of 96 individuals of Geum from the 39 sites sampled within the sympatric Berwickshire population. The mean number of SNPs successfully scored per individual was 207.3 (range 161 – 215 Figure S5). The variation in hybrid index (HI) in the Berwickshire population is illustrated in Figure 3.

The vast majority of individuals in the total sympatric sample (92 out of 96) have genotypes in which more than 97% of the species-specific SNPs are derived from only one of the two Geum species. Individuals possessing >97% G. rivale variants are hereafter regarded as G. rivale, while plants with
>97% *G. urbanum* variants are classified as *G. urbanum*. The four remaining individuals with substantial proportions of variants from both species correspond to the four early generation hybrids identified in the fastSTRUCTURE clustering analysis. The correlation between the Q values calculated by fastStructure and the hybrid index calculated here is high (r=0.99) (see Figure S6 for comparison using 188 SNPs), in part because some SNPs (i.e., 61 of the 492 SNPs available for fastSTRUCTURE analysis) are common to both analyses. Early generation hybrids were significantly associated with ‘mixed’ rather than ‘pure’ sites within Berwickshire (Fisher’s exact test, P=0.044).

Neither of the two individuals possessing a hybrid index close to 0.5 are heterozygous at all SNPs, the situation anticipated if they were simple F1 hybrids (Table 3). Instead they are homozygous at a substantial fraction of the species-specific SNPs (15% and 18%), suggesting that they may have been derived by selfing of F1 hybrids. Two further individuals contain alleles derived predominantly from *G. rivale*, but additionally possess a substantial complement of *G. urbanum* alleles (29% and 12% respectively) (Table 5). On the basis of their complement of alleles alone, these plants are most likely to represent first and second generation backcrosses to *G. rivale* respectively. However, their origin must again be more complex, possibly involving selfing, because a substantial fraction of the *G. urbanum* alleles they possess (10% and 33% respectively) are present in homozygous form, whereas if they had been simple backcrosses, all *G. urbanum* alleles would have been present as heterozygotes.

Among the 92 plants from Berwickshire possessing a preponderance (>97%) of species-specific alleles from one taxon, 33 of the 47 individuals assigned to *G. rivale* and 26 of the 45 plants assigned to *G. urbanum* possessed from 1 to 10 alleles classified as specific to the alternate species (Fig. S7). These could represent alleles that have introgressed, or alternatively alleles present at low frequency in the focal species that have not been detected in the limited sample of allopatric genotypes used to identify species-specific SNPs. The maximum frequencies of putatively introgressed alleles are low.
(9.6% and 14.4% in *G. rivale* and *G. urbanum* respectively). Furthermore 76% are polymorphic in *G. rivale* only, 20% are polymorphic in *G. urbanum* only, and 4% are polymorphic in both species. These proportions are not significantly different from those found in the total sample of SNPs scored (P>0.05, Table 1). Therefore, misclassification of species-specific SNPs rather than introgression is the most likely explanation for the apparent presence of up to 3% admixture in the *Geum* genomes from sympatric populations.

With one exception, the putatively introgressed alleles discussed above were randomly associated with each other within each of the sympatric *G. rivale* and *G. urbanum* populations. However, two SNPs at radtags 20454 and 24791 respectively, and originally classified as specific to *G. rivale*, showed complete association in genotypic state within sympatric *G. urbanum*. The alleles were present at a mean frequency of 14.4% and distributed across six sites. In six individuals both alleles were present in homozygous form, while a seventh individual was heterozygous at both radtags.

Since the SNPs involved are located on different scaffolds, these results may indicate the presence of a section of genome spanning the two scaffolds, possibly derived from *G. rivale*, which is segregating in the *G. urbanum* population.

**Analysis of Morphological Variation in Sympatric Populations**

**i. Correspondence between Phenotypic and Genotypic classification**

Morphological measurements of eight floral and vegetative characters commonly used to distinguish *G. rivale* and *G. urbanum* were obtained from 87 of the 96 individuals from Berwickshire that had been scored for species-specific SNPs. The results of Principal Component Analysis based on these morphological data are illustrated in Figure 4. The first two principal components, accounting for 82.4% and 6.7% of the variation respectively, distinguish two major groups, with two individuals falling between these on the first PCA axis. Individuals classified as *G. rivale* (N=43) and *G. urbanum* (N=40) on the basis of species-specific SNPs fall clearly into the two major groups with +ve and –ve
PCA1 values, respectively. The two individuals with intermediate PCA1 scores correspond to the early
generation hybrids identified with clustering and species-specific SNPs. The two genotypes identified
as complex backcrosses to G. rivale, on the basis of species-specific SNPs, group morphologically with
individuals classified as G. rivale on the PCA plot.

**ii. Morphological Variation and Putative Introgression**

To determine whether presence of putatively introgressed genetic material marked jointly by alleles
at radtags 20454 and 24792 affected the morphology of G. urbanum (see above), data on the eight
floral and vegetative characters measured in the sympatric G. urbanum population were used to
generate the first two principal component scores describing morphological variation. Analysis of
variance showed no significant effect of joint presence/absence of the alleles on PCA1, but a
significant effect on PCA2 score ($F_{1,38} = 8.67, P=0.005$; Fig.5). Individuals of G. urbanum jointly
possessing these alleles showed a significantly larger flower angle than those lacking them (91.4 vs.
60.8 degrees, $p=0.018$). Greater flower angle is a feature characteristic of G. rivale and is most
heavily weighted in PCA2.
DISCUSSION

We have developed a set of sequence based markers in two closely related non-model plant species with contrasting mating systems that share a polyploid history. Analysis of the markers across allopatric, and pure and mixed sympatric samples of *G. rivale* and *G. urbanum* indicate that the species are 2-3MY old, and that levels of introgression between the species, both before and after the period of recent sympatry, have been very small. While hybrid genotypes continue to be produced in contemporary populations, the proportion of introgressed loci across the genomes of the vast majority of individuals is much lower than 3%.

The development of sequence based markers, using next generation sequencing, poses technical problems in our *Geum* species. Their genomes are large (1.2 and 1.6 Gb in *G. urbanum* and *G. rivale* respectively) and although inheritance is disomic the species are ancient hexaploids. This means that there will generally be three closely related copies of every sequence, creating significant difficulties in distinguishing homologs from paralogs (Smedmark *et al.* 2003). We adopted the ddRAD technique that allows modulation of the number of RAD tags generated per genome to ensure adequate depth of coverage (Peterson *et al.* 2012). To maximize the chance of distinguishing homologs and paralogs we used paired end sequencing yielding long reads, and generated a draft genome of the largely homozygous inbreeding species against which to map our ddRADtag sequences. During the data analysis phase we employed very rigorous filtering of scaffolds and SNPs to exclude paralogs, which removed about three-quarters of scaffolds from our dataset. This filtering contributed to the development of a modest number of SNPs compared to comparable studies in other genera (e.g. *Populus*, Christe *et al.* 2016; *Betula*, Zohren *et al.* 2016), but that was nevertheless orders of magnitude higher than for previous microsatellite based studies; moreover, we developed sufficient SNPs to yield high precision estimates of introgression (e.g., 200 SNPs can estimate introgression of...
Furthermore our RAD data gave estimates of both inbreeding coefficients and population genetic differentiation that are congruent with results from microsatellite markers, providing confidence in the conclusions drawn from their analysis (Ruhsam et al. 2010, 2013).

Coalescent analyses of RAD data from allopatric populations, which contrast the fit of simple models of the species’ histories to polymorphism data, gave species’ age estimates of between 2 and 3 million years. This is considerably older than some estimates of time of divergence for other sister taxa that possess contrasting outcrossing and selfing mating systems. These range from 50 - 100kya for Capsella grandiflora and C. rubella (Brandvain et al. 2013) through 500-1100kya for Mimulus guttatus and M. nasutus (Aeschbacher et al. 2016) to 1Mya for Arabidopsis lyrata and A. thaliana (Tang et al. 2007). Our estimated age of the Geum species is consistent with the numerous morphological and physiological differences that are exhibited between them (Taylor 1997a, b).

However despite their age, G. rivale and G. urbanum apparently exhibit few intrinsic hybrid incompatibilities (Ruhsam et al. 2013); these observations parallel those from Drosophila species, which show few post-zygotic incompatibilities for species of a similar age (Coyne & Orr 1989, 1997).

Coalescent analyses also show that low levels of effective introgression are likely to have occurred between the Geum species up to the time of isolation of allopatric populations, but there is no strong signal of introgression in a particular direction. For clarity, M measures the long-term effective migration, and accounts for reduced gene migration due to selection against hybrids (see below).

Our estimate of the introgression rate between the Geum species (M=0.04 diploid genomes per generation) is an order of magnitude lower than that estimated for ancient introgression from inbreeding M. nasutus into outcrossing M. guttatus (M=0.1-1; Aeschbacher et al. 2016) and between outcrossing and selfing subspecies of Clarkia in their secondary contact zone (M=0.897-0.169; Pettergill & Moeller 2012). It is many orders of magnitude below contemporary estimates of
unidirectional gene migration from the inbreeding rice Oryza nivaria to its outcrossing relative O. rufipogon (M=2.3 \times 10^5; Zhou et al. 2008).

Our coalescent analyses necessarily rely on simplifying assumptions. First, the blockwise site-frequency spectrum approach we adopted assumes a constant \( N_e \) and \( \mu \) across blocks and so ignores effects of heterogeneity in mutation rate and the effects of background selection on linked sites, which could lead to a spurious signal of introgression (Ewing & Jensen 2016). Second, our models are drastic simplifications of the history of these species and, in particular, assume constant \( N_e \) in G. rivale and the ancestral population, which is unrealistic given the climatic fluctuations and the associated range shifts of European taxa during the Pleistocene. Following species divergence, either a change in \( N_e \) or introgression can alter the joint site-frequency spectrum (Chen 2012); by assuming \( N_{\text{anc}} = N_{\text{riv}} \), we potentially conflate changes in \( N_e \) with introgression, which, again, could lead to a spurious signal of introgression. Given these issues and the low estimates of introgression, we do not interpret the support of IM models over the div\(_2\) model as strong evidence for ancient long-term gene flow between the Geum species.

As a first test of more recent introgression in Geum we conducted comparisons of genetic diversity and differentiation between species in allopatry vs. sympatry. Introgression, which is only possible in sympatric populations, can potentially increase genetic diversity of the recipient species, and erode genetic differences between species. Hence, our comparisons of \( \pi \) and dxy between sympatric and allopatric populations can provide evidence for or against introgression. Relative genetic diversity was greater for G. rivale than G. urbanum, as expected based on their different mating systems (Charlesworth 2003). In addition, genetic diversity was approximately 9% higher for G. rivale in sympathy than in allopatry, but was equal between population types for G. urbanum. While this is consistent with introgression into G. rivale, but not G. urbanum, other factors could account for this
result, including a greater effective population size of *G. rivale* in sympatry than in allopatry.

Furthermore, genetic differentiation between species (dxy) was similar in sympatry and allopatry, providing no evidence for introgression.

As a second test of recent introgression we analysed the genomic composition of individuals from a region of sympatry in Berwickshire and compared them with samples from isolated, reference populations. The cluster analysis performed by fastSTRUCTURE (which did not rely exclusively on species-diagnostic SNPs), and the analysis based on a limited subset of species-diagnostic SNPs identified the same four individuals as early hybrids and backcrosses, and yielded similar genomic composition estimates that were congruent with morphological classifications. Critically, they suggest that the upper limit for the proportion of introgressed genome present in Berwickshire plants is ~3%. However, the real value for introgression is likely to be lower because the assumption that all our species diagnostic SNPs are alternately fixed between the species is unlikely to be true.

Although our overall results indicate very little recent genetic exchange, one intriguing result provides tentative evidence for introgression from *G. rivale* to *G. urbanum*. Two SNP variants located on different scaffolds that were classified as specific to *G. rivale* in allopatry show co-segregation within the sympatric *G. urbanum* population. Moreover, *G. urbanum* individuals possessing these alleles show a significantly greater flower angle, a characteristic of *G. rivale*. A possible explanation is introgression of a chromosome segment from *G. rivale* into *G. urbanum* that is now polymorphic within the sympatric population. Given the absence of evidence for backcrossing of hybrids to *G. urbanum* (Ruhsam *et al.* 2011, 2013), such introgression could only arise through the establishment, following hybridisation, of a self-fertilising lineage that largely retains genetic material from *G. urbanum*. Previous work has shown that F1 hybrids are self-compatible and a proportion may be capable of autopollination, so this scenario, though unlikely, is not impossible (Ruhsam *et al.* 2013).
Despite the reproductive barrier that adoption of a selfing mating system naturally imposes (Martin & Willis 2007; Briscoe Runquist et al. 2014; Brys et al. 2014), the widespread occurrence of the F1 hybrid G. intermedium throughout the area of sympatry of G. rivale and G. urbanum in Berwickshire and elsewhere (Preston et al. 2002), and our own detection of early generation hybrids, indicates that low levels of introgression are not caused by lack of opportunity for hybridisation. Previous studies have shown that where hybridisation occurs, a wide array of fertile recombinant offspring are produced both by backcrossing with G. rivale and by self-fertilisation and that under benign conditions these show little evidence of intrinsic genetic incompatibility (Ruhsam et al. 2011, 2013).

Thus G. rivale and G. urbanum join a growing list of sister species which readily produce F1 hybrids and viable recombinant offspring, yet show little or no signal of introgression (e.g. Encelia (Kyhos et al. 1981); Quercus, Arctostaphylos, Iris (Nason et al. 1992); Rhododendron (Milne et al. 2003); Begonia (Twyford et al. 2015); Populus (Christe et al. 2016); Bombina (Nürnberg et al. 2016).

In many of the above examples ecological selection against recombinants has been invoked to explain the maintenance of species distinctions. There are good grounds for proposing the same explanation to account for lack of introgression in Geum. Since their initial divergence the Geum species have evolved marked differences in pollination syndrome, phenology, vegetative morphology and many aspects of physiology including tolerance of shade and waterlogging (Waldren et al. 1988; Taylor 1997a, b). Therefore multiple ecologically important characters could potentially be the targets of selection against recombinant individuals in natural populations (Nosil et al. 2009).

Selection may be particularly effective in the self-fertilising recombinants that are required to allow introgression from G. rivale into G. urbanum. Here reduced effective recombination rates will lead to correlated selective effects across the genome, and deleterious recessive mutations present as genetic load in genome segments derived from outcrossing G. rivale will be exposed (Hu 2015). Given
our ability readily to artificially cross the *Geum* species, and produce large numbers of later generation hybrids, there is now the potential experimentally to test the hypothesis that strong ecological selection maintains species integrity using appropriate field experiments.
Author contributions

C.Y.J, K.G. and R.A.E. conceived the study; C.Y.J., F.T., M.T. and R.A.E. gathered data; C.Y.J., K.L., F.T. and R.A.E. analysed data; C.Y. J., K.L. and R.A.E. wrote the manuscript with input and revisions from all co-authors.

Acknowledgements: This research was funded by BBSRC grant Bb/J006580/1 to RAE. We are very grateful to Barbara Giles, Cecile Bacles, Olivier Lepais, Hjörtur Þorbjörnsson, and Jorge Paiva who provided non-UK samples for this project. Beate Nürnberg and Alex Tywford provided comments that improved this paper.

Data accessibility

DNA sequences: will be made available on European Nucleotide Archive before publication

Assembled genome: will be made available (University of Edinburgh link) before publication

Morphological data: will be made available on Dryad before publication

Scripts: will be made available on github.com before publication

Conflict of interest: None declared.
REFERENCES

Abbott RJ, Barton NH, Good JM (2016) Genomics of hybridisation and its evolutionary consequences. *Molecular Ecology*, 25, 2325-2332.

Arnold B, Corbett-Detig RB, Hartl D, Bomblies K. (2013) RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. *Molecular ecology*, 22, 3179-3190.

Aeschbacher S, Selby JP, Willis JH, Coop, G (2016) Population-genomic inference of the strength and timing of selection against gene flow. doi: https://doi.org/10.1101/072736.

Arens P, Durka W, Wernke-Lenting JH, Smulders MJM (2004) Isolation and characterization of microsatellite loci in *Geum urbanum* (Rosaceae) and their transferability within the genus *Geum*. *Molecular Ecology Notes*, 4, 209–212.

Arunkumar R, Ness RW, Wright SI, Barrett SCH (2015) The evolution of selfing is accompanied by reduced efficacy of selection and purging of deleterious mutations. *Genetics*, 199, 817-29.

Barrett SCH, R. Arunkumar, R, Wright SI (2014) The demography and population genomics of evolutionary transitions to self-fertilization in plants. *Philosophical Transactions of the Royal Society B*, 369, 2013034.

Brandvain Y, Slotte T, Hazzori KM, Wright SI, Coop G (2013) Genomic identification of founding haplotypes reveals the history of the selfing species *Capsella rubella*. *PLoS Genetics*, 9, e1003754.

Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL (2014) Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. *PLoS Genetics*, 10, e1004410.

Briscoe Runquist RD, Chu E, Iverson JL, Kopp JC, Moeller DA (2014) Rapid evolution of reproductive isolation between incipient outcrossing and selfing *Clarkia* species. *Evolution*, 68, 2885-2900.

Brys R, Vanden Broeck A, Mergeay J, Jacquemyn H (2014) The contribution of mating system variation to reproductive isolation in two closely related *Centaurium* species (Gentianaceae) with a generalized flower morphology. *Evolution*, 68, 1281–1293.

Catchen JM, Hohenlohe PA, Bassham S, Amores A, Cresko WA (2013) Stacks: an analysis tool set for population genomics. *Molecular Ecology*, 22, 3124-3140

Catchen JM, Amores A, Hohenlohe P, Cresko W, Postlethwait JH (2011) Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, 1, 171-182.

Carney SE, Gardner KA, Rieseberg LH (2000) Evolutionary changes over the fifty-year history of a hybrid population of sunflowers (*Helianthus*). *Evolution*, 54, 462–474.

Charlesworth D (2003) Effects of inbreeding on the genetic diversity of populations. *Philosophical Transactions of the Royal Society B*, 358, 1051–1070.

Charlesworth D, Morgan MT, Charlesworth B (1993) Mutation accumulation in finite outbreeding and inbreeding populations. *Genetics Research*, 61, 39-56.

Charlesworth D, Wright SI (2001) Breeding systems and genome evolution. *Current Opinion in Genetics & Development*, 11, 685-690.

Chen H (2012) The joint allele frequency spectrum of multiple populations: A coalescent theory approach. *Theoretical Population Biology*, 81, 179-195.
Christe C, Stolting KN, Bresadola L, Fussi B, Heinze B, Wegmann D, Lexer C (2016) Selection against recombinant hybrids maintains reproductive isolation in hybridizing *Populus* species despite F1 fertility and recurrent gene flow. *Molecular Ecology*, **25**, 2482-2498.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133-3157.

Coyne JA, Orr HA (1989). Patterns of speciation in Drosophila. *Evolution*, **43**, 362-381.

Coyne JA, Orr HA (1997) "Patterns of speciation in Drosophila" revisited. *Evolution*, **51**, pp.295-303.

Coyne JA, Orr H A (2004) *Speciation*. Sinauer Associates, Sunderland, Massachusetts.

Doyle JJ, Doyle JL (1990) Isolation of plant DNA from fresh tissue. *Focus*, **12**, 13–15.

Ducarmé V, Vrancken J, Wesselingh RA (2010) Hybridization in annual plants: patterns and dynamics during a four-year study in mixed Rhinanthus populations. *Folia Geobotanica*, **45**, 387-405.

Ewing GB, Jensen JB (2016) The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, **25**, 135-141.

Fishman L, Wyatt R (1999) Pollinator-mediated competition, reproductive character displacement, and the evolution of selfing in *Arenaria uniflora* (Caryophyllaceae). *Evolution*, **53**, 1723-1733.

Glénin S, Ronfort J (2013) Adaptation and maladaptation in selfing and outcrossing species: New mutations versus standing variation. *Evolution*, **67**, 225-240.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genetics*, **6**, e1000862.

Hu X-S (2015) Mating system as a barrier to gene flow. *Evolution*, **69**, 1158-1177.

Igic B, Bohs L, Kohn JR (2006) Ancient polymorphism reveals unidirectional breeding system shifts. *Proceedings of the National Academy of Sciences*, **103**, 1359-1363.

Kenny AM, Sweigart AL (2016) Reproductive isolation and introgression between sympatric *Mimulus* species. *Molecular Ecology*, **25**, 2499–2517

Kyhos DW, Clark C, Thompson WC (1981) The hybrid nature of *Encelia laciniata* (Compositae: *Heliantheae*) and control of population composition by post-dispersal selection. *Systematic Botany*, **6**, 399–411.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler Transform. *Bioinformatics*, **25**, 1754-60.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, **25**, 2078-2079.

Lohse K, Chmelik M, Martin SH, Barton NH (2016) Efficient strategies for calculating blockwise likelihoods under the coalescent. *Genetics*, **202**, 775-786.
Martin NH, Willis JH (2007) Ecological divergence associated with mating system causes nearly complete reproductive isolation between sympatric Mimulus species. Evolution, 61, 68–82.

Milne, R.I., Terzioglu, S. & Abbott, R. J. (2003). A hybrid zone dominated by fertile F1s: maintenance of species barriers in Rhododendron. Molecular Ecology, 12, 2719-2729.

Nason JD, Ellstrand NC, Arnold ML (1992) Patterns of hybridization and introgression in populations of oaks, manzanitas, and irises. American Journal of Botany, 79, 101-111.

Nei M, Li WH (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proceedings of the National Academy of Sciences, 76, 5269-5273.

Nosil P, Harmon LJ, Seehausen O (2009) Ecological explanations for (incomplete) speciation. Trends in Ecology & Evolution, 24, 145-156.

Nürnberg B, Lohse K, Fijarczyk A, Szymura JM, Blaxter ML (2016) Para-allopatry in hybridising fire-bellied toads (Bombina bombina and B. variegata): Inference from transcriptome-wide coalescence analyses. Evolution, 70, 1803-1818.

Ossowski S, Schneeberger K, Lucas-Ledó JI, Warthmann N, Clark RM, Shaw, RG, Weigel D, Lynch, M (2010) The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. Science, 327, 92-94.

Paolletti M, Buck KW, Brasier CM (2006) Selective acquisition of novel mating type and vegetative incompatibility genes via interspecies gene transfer in the globally invading eukaryote Ophiostoma novo-ulmi. Molecular Ecology, 15, 249–262.

Parra G, Bradnam K, Korf I (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics, 23, 1061–1067.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one, 7, e37135.

Pettergill JB, Moeller DA (2012) Phylogeography of speciation: allopatric divergence and secondary contact between outcrossing and selfing Clarkia. Molecular Ecology, 21, 4578-4592.

Preston CD, Pearman DA, Dines TD (2002) New Atlas of the British and Irish flora. Oxford University Press, Oxford.

Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics, 197, 573-589.

Ruhsam M, Hollingsworth PM, Squirrel J, Ennos RA (2010) Significant differences in outcrossing rate, self-incompatibility, and inbreeding depression between two widely hybridizing species of Geum. Biological Journal of the Linnean Society, 101, 977–990.

Ruhsam M, Hollingsworth PM, Ennos RA (2011) Early evolution in a hybrid swarm between outcrossing and selfing lineages in Geum. Heredity, 107,246–255.

Ruhsam M, Hollingsworth PM, Ennos RA (2013). Patterns of mating, generation of diversity, and fitness of offspring in a Geum hybrid swarm. Evolution, 67, 2728–2740.

Sicard A, Lenhard M (2011) The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. Annals of Botany, 107, 1433-1443.
Stebbins GL (1957) Self-fertilization and population variability in the higher plants. *American Naturalist, 91*, 337-354.

Tang C, Toomajian C, Sherman-Broyles S, Plagnol V, Guo Y-L, Hu TT, Clark RM, Nasrallah JB, Weigel D, Nordborg M (2007). The evolution of selfing in *Arabidopsis thaliana*. *Science, 317*, 1070-1072.

Taylor K (1997a) Biological flora of the British Isles: *Geum rivale*. *Journal of Ecology, 85*, 721–731.

Taylor K (1997b) Biological flora of the British Isles: *Geum urbanum*. *Journal of Ecology, 85*, 705–720.

Twyford AD, Kidner CA, Ennos RA (2015) Maintenance of species boundaries in a Neotropical radiation of *Begonia*. *Molecular Ecology, 24*, 4982-4993.

Vandepitte K, Jacquemyn H, Roldan-Ruiz, I, Honnay O (2007) Landscape genetics of the self-compatible forest herb *Geum urbanum*: effects of habitat age, fragmentation and local environment. *Molecular Ecology, 16*, 4171-4179.

Vickery RK Jr (1978) Case studies in the evolution of species complexes in *Mimulus*. *Evolutionary Biology, 11*, 405–507.

Waldren S, Etherington JR, Davies MS (1988) Comparative studies of plant growth and distribution in relation to waterlogging XV. The effect of waterlogging on growth of various populations of and hybrids between *Geum rivale* L. and *Geum urbanum* L. *New Phytologist, 109*, 97 – 106.

Waldren S, Etherington JR, Davies MS. (1989) Hybridization and ecology of *Geum rivale* L and *Geum urbanum* L (*Rosaceae*) at Malham Tarn, Yorkshire. *Botanical Journal of the Linnean Society, 101*, 199–211.

Wright SI, Ness RW, Foxe JP, Barrett SCH (2008) Genomic consequences of outcrossing and selfing in plants. *International Journal of Plant Sciences, 169*, 105-118.

Wu TD, Watanabe CK (2005) Gmap: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics, 21*, 1859–1875.

Wright SI, Kalisz S, Slotte T (2013) Evolutionary consequences of self-fertilization in plants. *Proceedings of the Royal Society B, 280*, 20130133.

Zhou H-F, Zheng X-M, Wei R-X, Second G, Vaughan DA, Ge S (2008). Contrasting population genetic structure and gene flow between *Oryza rufipogon* and *Oryza nivara*. *Theoretical and Applied Genetics, 117*, 1181–1189.

Zohren J, Wang N, Kardailsky I, Borrell JS, Joecker A, Nichols RA, Buggs RJA (2016) Diploid–tetraploid introgression among British birch trees with shifting ranges shown by restriction site-associated markers. *Molecular Ecology, 25*, 2413–2426.
| Polymorphism Type                                                                 | Count |
|----------------------------------------------------------------------------------|-------|
| Alternately fixed SNPs                                                           | 488   |
| SNP polymorphic in *G. rivale* but one allele fixed in *G. urbanum* (e.g., *G. rivale*: A, T; *G. urbanum*: T) | 1334  |
| SNP polymorphic in *G. urbanum* but one allele fixed in *G. rivale*              | 338   |
| *G. rivale* and *G. urbanum* share polymorphism                                  | 34    |
| Total SNPs analysed                                                               | 2194  |

Table 1: Frequency of polymorphism of different forms within and between *Geum* species. Analysis allows multiple SNPs per scaffold.
Table 2. $F_{IS}$ estimates for ‘allopatric’ UK *G. rivale* and *G. urbanum* populations using data filtered for paralogs.
| Sample          | div₂ | IM₀₂ | IM₀₄ |
|-----------------|------|------|------|
| Ben Lui 1       | -12.9| -12.5| 0    |
| Ben Lui 4       | -9.66| 0    | -0.721|
| Ben Lawers 5    | -6.67| -6.51| 0    |

Table 3. Difference in log Likelihoods between the “best” and alternate models. “Sample” refers to which *G. rivale* individual is paired with the single *G. urbanum* individual.
Table 4. Parameter estimates under the models for *G. rivale* sample, Ben Lui 4. 95% Confidence intervals provided in parentheses. See Table S6 for parameter estimates for all three sample pairs.

| Model        | $\text{div}_2$ | $\text{IM}_{\text{rul}}$ | $\text{IM}_{\text{ur}}$ |
|--------------|----------------|---------------------------|--------------------------|
| $N_{\text{anc}} = N_{\text{rivb}}$ | $1.30 \times 10^5$ | $1.24 \times 10^5$ | $1.11 \times 10^5$ |
| $N_{\text{urb}}$ | $3.77 \times 10^4$ | $3.09 \times 10^4$ | $3.38 \times 10^4$ |
| t (years)    | $2.22 \times 10^6 (2.01 \times 10^6 - 2.43 \times 10^6)$ | $2.35 \times 10^6$ | $2.48 \times 10^6$ |
| $M$ (mean, 95% CI’s) | 0.04 (0.005 - 0.191) | 0.04 (0.007 - 0.101) |
| Sample and classification | % ‘rivale’ in genome | Homozygous ‘rivale’ allele | Heterozygous | Homozygous ‘urbanum’ allele | Total Loci scored |
|---------------------------|-----------------------|---------------------------|--------------|---------------------------|------------------|
| 29A (H)                   | 47.2                  | 9                         | 166          | 20                        | 195              |
| 32C (H)                   | 45.3                  | 7                         | 132          | 22                        | 161              |
| 26B (BCR)                 | 87.9                  | 138                       | 28           | 7                         | 173              |
| 27B (BCR)                 | 70.8                  | 95                        | 113          | 6                         | 214              |

Table 5 Genotypic composition at species-specific ddRAD loci for individuals classified as early generation hybrids (H) and backcrosses to *G. rivale* (BCR).
Figure 1. Locations of allopatric populations (large map) of *Geum rivale* and *G. urbanum* and detailed distribution of 39 sympatric populations in eight 10km x 10km squares (inset).
Figure 2. Results of fastSTRUCTURE (Raj et al. 2014) cluster analysis for K=2 based on 188 SNP loci present in all 133 Geum individuals scored in the allopatric and sympatric populations. Each bar represents one individual and shows the proportion of the genome from G. rivale (red) and G. urbanum (yellow).
Figure 3. Distribution of hybrid index (proportion of *G. rivale* species-specific alleles) for 96 individuals within 39 sympatric populations where *G. rivale* and *G. urbanum* were either found alone (blue bars, ‘pure population’) or together (green bars, ‘mixed population’).
Figure 4. PCA analysis based on eight morphological characters measured in a common environment for a sample of 87 individuals of Geum sampled from sympatric populations. Genetic class (from Hybrid Index) of each individual (G. rivale, G. urbanum, early hybrid, or backcross to G. rivale) derived from an analysis of 215 species-specific markers is shown.
Figure 5. PCA analysis of 40 individuals of *G. urbanum* sampled from sympatric populations based on variation for eight morphological characters measured in a common environment. Red, square individuals jointly possess species-specific SNPs from *G. rivale* at radtags 20454 and 24791.
SUPPLEMENTARY MATERIAL

i. Development of Draft Genome for G. urbanum

Reads were generated from three Illumina libraries: a 151 paired end library with a mean insert size of 250 (sequenced on Hi seq 2500 (v1 chemistry) to give 144 million read pairs), a 151 paired end library with a mean insert size of 450 (sequenced on Hi seq to give 134 million read pairs), and a 251 paired end library with a mean insert size of 450 (sequenced on Illumina Mi seq (v2 chemistry) to give 11.9 million read pairs). Illumina adapters were removed and reads quality trimmed using cutadapt version 1.3 (Marcel, 2011) with the option “-q 30”. Trimmed reads were assembled into contigs and scaffolds using MaSuRCA version 2.0.3.1; MaSuRCA (Zimin et al., 2013) was run using the command line options:

```
GRAPH_KMER_SIZE=auto, USE_LINKING_MATES=1, LIMIT_JUMP_COVERAGE = 60, ovlMemory=30,
cgwErrorRate=0.25, ovlMemory=4GB, KMER_COUNT_THRESHOLD = 1, NUM_THREADS=
64,JF_SIZE=23000000000, DO_HOMOPOLYMER_TRIM=0.
```

For each individual sample, 35ul of genomic DNA (250ng) was digested in a total volume of 50 ul, comprising 1.0 ul Sphl-HF (20u), 1.0 ul EcoRI-HF (20u), 5.0 ul 10x Cutsmart buffer (New England Biolabs, UK) and 8.0 ul dH2O for a total of 3 hours, at 37°C. The reaction was purified by the addition of Ampure XP beads at a ratio 1.8 x beads: 1x DNA rather than by heat denaturation, to maintain the integrity of the EcoRI generated overhangs.

Samples were eluted in 25ul EB (Qiagen). Unique barcoded EcoRI P1 adapters (Peterson et al. 2012) and a generic SphI P2 adapter (Table S3) were ligated to the complementary overhangs of the DNA fragments generated by the EcoRI:SphI double digest reaction. Each ligation reaction was carried out in a total volume of 40ul, consisting of 25ul DNA, 2.0 ul P1 adapter (4uM), 2.0 ul P2 adapter (4uM), 4.0 ul 10x T4 Ligase buffer (New England Biolabs, UK) 1ul Quick Ligase (New England Biolabs, UK), and the reaction was incubated at 23°C for 30 min followed by heat denaturation at 65 °C for 15 min.

After ligation the reactions were pooled into libraries containing a maximum of 24 samples, each barcoded with a unique EcoRI P1 adapter sequence. The pooled libraries were then subjected to two rounds of Ampure XP bead clean up, at a ratio of 1x Ampure XP beads: 1 x DNA and eluted in a volume of 30ul EB (Qiagen). Individual libraries were size selected by BluePippin Size selection system, 2% Precast Agarose gel cassette (sage science) using the ‘Tight’ collection mode set to a target size of 300bp. The success of size selection was assessed by Bioanalyzer DNA HS chip (Agilent Technologies).

For each ddRAD library, PCR amplification was carried out in 2 independent 50 μL reactions consisting of 20 ul size selected DNA, 25ul 2x Phusion Master Mix (New England Biolabs, UK), 2.5 ul DMSO, and 1ul ddRAD P1 PCR primer (10 nm), and a barcoded P2 primer (10 nm) that introduced a 8bp unique index during amplification (Table S3), using the following cycling parameters: 98 °C for 30 s followed by 12 cycles of 98 °C for 10 s and 72 °C for 60 s.

PCR products were then pooled, and purified by Ampure XP beads at a ratio of 0.8 x beads: 1x DNA, and the mean insert size of the individual libraries was assessed by Bioanalyzer DNA HS chip, (Agilent Technologies), and were quantified by qPCR on an Illumina Eco instrument using the Kapa Library Quantification Kit.mA final super pool was created consisting of equimolar individual library PCR reactions and the quantification of the pool was measured by qPCR.
References

Marcel M (2012). Cutadapt removes adapter sequences from high-throughput sequencing reads. Bioinformatics in Action, 17, 10-12.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PloS one, 7, e37135.

Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics, 29, 2669-2677.
### SUPPLEMENTARY TABLES

| Site                  | Coordinates                  | Taxon |
|-----------------------|------------------------------|-------|
| Ben Lawers, UK        | 56°32′23.46″N/ 004°17′27.58″W | R (4) |
| Coire Garblach, UK    | 57°01′38.29″N/ 003°50′08.17″W | R (4) |
| Ben Lui, UK           | 56°23′38.87″N/ 004°49′28.02″W | R (4) |
| Iceland               | 64°07′N / 21°50′W          | R (1) |
| Sweden                | 63°17′27.00″N/ 18°42′54.00″E | R (1) |
| Selwyn Wood, UK       | 50°57′43.96″N/ 000°12′40.52″E | U (2) |
| Combe Wood, UK        | 51°02′05.03″N/ 000°18′31.39″E | U (2) |
| Burgh Wood, UK        | 51°01′21.30″N/ 000°27′13.26″E | U (2) |
| Hoads Wood, UK        | 51°08′47.89″N/ 000°47′27.02″E | U (2) |
| Stanford Bridge, UK   | 51°09′00.56″N/ 000°45′33.55″E | U (2) |
| Frith Wood, UK        | 51°10′22.20″N/ 000°42′59.47″E | U (2) |
| Priory Wood, UK       | 51°05′07.33″N/ 000°54′02.91″E | U (2) |
| Copperhurst, UK       | 51°04′55.62″N/ 000°56′53.11″E | U (2) |
| Mill Wood, UK         | 50°58′47.12″N/ 000°24′56.60″E | U (2) |
| Punnets Town, UK      | 50°58′20.12″N/ 000°18′43.66″E | U (2) |
| Portugal              | 40°22′43.00″N/ 008°22′11.00″E | U (1) |
| France                | 43°16′59.57″N/ 001°28′56.59″E | U (1) |

Table S1. Sampling locations for allopatric populations of *Geum rivale* (R) and *G. urbanum* (U). Numbers in brackets = sample size.
| Site Number | Elevation(m) | Coordinates                        | Taxa |
|------------|-------------|------------------------------------|------|
| 1          | 331         | 55°49'26.54"N/ 002°46'46.78"W      | R (2)|
| 2          | 271         | 55°46'34.73"N/ 002°47'05.62"W      | U (2)|
| 3          | 360         | 55°47'24.57"N/ 002°43'31.04"W      | R (2)|
| 4          | 300         | 55°42'00.02"N/ 002°45'29.21"W      | U (2)|
| 5          | 232         | 55°44'28.64"N/ 002°38'33.57"W      | R (2)|
| 6          | 180         | 55°39'10.43"N/ 002°41'50.99"W      | R (2) + U (2) |
| 7          | 100         | 55°41'36.40"N/ 002°42'32.59"W      | U (2)|
| 8          | 200         | 55°43'09.71"N/ 002°41'37.32"W      | R (2)|
| 9          | 230         | 55°44'09.73"N/ 002°40'25.34"W      | R (2)|
| 10         | 285         | 55°44'42.61"N/ 002°38'33.91"W      | R (2) + U (2) |
| 11         | 195         | 55°43'59.74"N/ 002°46'10.80"W      | U (2)|
| 12         | 230         | 55°45'01.53"N/ 002°33'59.10"W      | U (2)|
| 13         | 207         | 55°45'32.21"N/ 002°28'43.98"W      | R (2)|
| 14         | 300         | 55°48'48.07"N/ 002°35'30.28"W      | R (2)|
| 15         | 208         | 55°48'23.28"N/ 002°30'06.09"W      | R (2) + U (2) |
| 16         | 225         | 55°40'44.39"N/ 002°37'27.14"W      | R (2) + U (2) |
| 17         | 140         | 55°39'59.73"N/ 002°33'15.31"W      | R (2)|
| 18         | 175         | 55°41'09.26"N/ 002°33'42.29"W      | U (2)|
| 19         | 180         | 55°42'51.82"N/ 002°32'54.37"W      | R (2)|
| 20         | 190         | 55°42'05.55"N/ 002°35'57.70"W      | U (2)|
| 21         | 247         | 55°52'05.45"N/ 002°25'54.64"W      | R (2)|
| 22         | 213         | 55°49'57.38"N/ 002°26'06.32"W      | R (2)|
| 23         | 180         | 55°50'21.93"N/ 002°22'53.91"W      | U (2)|
| 24         | 236         | 55°51'59.51"N/ 002°25'42.03"W      | R (2) + U (2) |
| 25         | 120         | 55°55'13.74"N/ 002°20'45.68"W      | U (2)|
| 26         | 180         | 55°45'10.70"N/ 002°24'30.61"W      | R (2) + U(2)|
| 27         | 100         | 55°45'38.64"N/ 002°21'06.79"W      | R (2) + U (2) |
| 28         | 150         | 55°47'19.70"N/ 002°21'12.24"W      | R (2) + U (2) |
| 29         | 90          | 55°48'12.97"N/ 002°20'30.05"W      | U (2)|
|   |    |   |     |     |     |
|---|----|----|-----|-----|-----|
| 30 | 135 | 55°41′34.10″N/002°25′04.03″W | R (2) |   |
| 31 | 140 | 55°42′46.67″N/002°23′30.15″W | U (2) |   |
| 32 | 100 | 55°42′41.00″N/002°21′59.39″W | R (2) + U (2) |   |
| 33 | 60  | 55°40′58.23″N/002°21′03.75″W | U (2) |   |
| 34 | 130 | 55°42′46.72″N/002°25′01.36″W | R (2) |   |
| 35 | 220 | 55°55′03.85″N/002°15′55.46″W | R (2) |   |
| 36 | 140 | 55°53′14.06″N/002°17′24.90″W | R (2) + U (2) |   |
| 37 | 130 | 55°52′48.31″N/002°12′25.08″W | U (2) |   |
| 38 | 150 | 55°55′13.92″N/002°11′29.87″W | R (2) |   |
| 39 | 150 | 55°54′23.37″N/002°10′22.08″W | U (2) |   |

Table S2. Sampling locations for pure and mixed populations of *Geum rivale* (R) and *G. rivale* (U) within an area of geographic sympatry in Scotland. Numbers in brackets = sample size.
| Oligo ID            | EcoRI Adapter P1.1_Oligo sequence | EcoRI Adapter P1.2_Oligo sequence |
|---------------------|----------------------------------|----------------------------------|
| ACACGACA_EcoRI_P1.Top | 5'-ACACTCTTTTCCCTACACGACGCTCTTTCCGCATCTACACGACA-3' | |
| ACACGACA_EcoRI_P1.Bottom | 5'-Phos-AATTTGTCCTGATCGGAAAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| ACGTAGCA_EcoRI_P1.Top | 5'-ACACTCTTTTCCCTACACGACGCTCTTTCCGCATCTACGTAGCA-3' | |
| ACGTAGCA_EcoRI_P1.Bottom | 5'-Phos-AATTTGCTACGATCGGAAAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| ACTGCTCA_EcoRI_P1.Top | 5'-ACACTCTTTTCTTACACGACGCTCTTTCCGCATCTACGTAGCTCA-3' | |
| ACTGCTCA_EcoRI_P1.Bottom | 5'-Phos-AATTTGACGTCATGCGAAAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| AGCTGTGA_EcoRI_P1.Top | 5'-ACACTCTTTTCTTACACGACGCTCTTTCCGCATCTACGTAGCTGA-3' | |
| AGCTGTGA_EcoRI_P1.Bottom | 5'-Phos-AATTTACACAGCTAGCAGGAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| AGTCACGA_EcoRI_P1.Top | 5'-ACACTCTTTTCTTACACGACGCTCTTTCCGCATCTACGTACACGA-3' | |
| AGTCACGA_EcoRI_P1.Bottom | 5'-Phos-AATTTACAGCTAGCAGGAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| ATATCATA_EcoRI_P1.Top | 5'-ACACTCTTTTCTTACACGACGCTCTTTCCGCATCTATATACATA-3' | |
| ATATCATA_EcoRI_P1.Bottom | 5'-Phos-AATTTATAGATAGCATGCGAAAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| CACACAGT_EcoRI_P1.Top | 5'-ACACTCTTTTCTTACACGACGCTCTTTCCGCATCTACACAGT-3' | |
| CACACAGT_EcoRI_P1.Bottom | 5'-Phos-AATTTACTCTTGAGATCGGAAAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| CATGATCA_EcoRI_P1.Top | 5'-ACACTCTTTTCTTACACGACGCTCTTTCCGCATCTCATCTCACTA-3' | |
| CATGATCA_EcoRI_P1.Bottom | 5'-Phos-AATTTGATCATGAGATCGGAAAAGCGTCGTGTAGGGAAAGAGAGTG-3' | |
| CGCGCATA_EcoRI_P1.Top | 5'-ACACTCTTTTCTTACACGACGCTCTTTCCGCATCTCGCGCATA-3' | |
| Sequence                        | Primer 1 (Top)                      | Primer 2 (Bottom)                      |
|--------------------------------|------------------------------------|---------------------------------------|
| CGCGCATA_EcoRI_P1.Bottom       | 5'-Phos-AATTATAGCAGGGAAGAGAGCTCGTGTAGGGAAAGAGTGT-3' |
| CTAGTGTC_EcoRI_P1.Top          | 5'-ACACTTTCTCCCTACACGAGCAGCTCTCCCGATCTCTATGTC-3'   |
| CTAGTGTC_EcoRI_P1.Bottom       | 5'-Phos-AATTGAGCGAGGATCCGGAAGACGCTCGTGTAGGGAAAGAGTGT-3' |
| CTCTGCTC_EcoRI_P1.Top          | 5'-ACACTCTTTCCCTACACGAGCAGCTCTCCCGATCTCTCTGCTC-3'   |
| CTCTGCTC_EcoRI_P1.Bottom       | 5'-Phos-AATTGAGCGAGGATCCGGAAGACGCTCGTGTAGGGAAAGAGTGT-3' |
| CTGATGCT_EcoRI_P1.Top          | 5'-ACACTCTTTCCCTACACGAGCAGCTCTCCCGATCTCTGATGCT-3'   |
| CTGATGCT_EcoRI_P1.Bottom       | 5'-Phos-AATTGAGCGAGGATCCGGAAGACGCTCGTGTAGGGAAAGAGTGT-3' |
| GACTGCAG_EcoRI_P1.Top          | 5'-ACACTCTTCTCCCTACACGAGCAGCTCTCCCGATCTGACTGCAG-3'   |
| GACTGCAG_EcoRI_P1.Bottom       | 5'-Phos-AATTCTCGAGCGATCGGAAGAGCGTCTCGTGTAGGGAAAGAGTGT-3' |
| GATCGTG_A_EcoRI_P1.Top         | 5'-ACACTCTTCTCCCTACACGAGCAGCTCTCCCGATCTGATCGTGA-3'   |
| GATCGTG_A_EcoRI_P1.Bottom      | 5'-Phos-AATTCTCGAGCGATCGGAAGAGCGTCTCGTGTAGGGAAAGAGTGT-3' |
| GCATGTGC_EcoRI_P1.Top          | 5'-ACACTCTTCTCCCTACACGAGCAGCTCTCCCGATCTGCATGTGC-3'   |
| GCATGTGC_EcoRI_P1.Bottom       | 5'-Phos-AATTCTCGAGCGATCGGAAGAGCGTCTCGTGTAGGGAAAGAGTGT-3' |
| GCTACAGC_EcoRI_P1.Top          | 5'-ACACTCTTCTCCCTACACGAGCAGCTCTCCCGATCTGCTACAGC-3'   |
| GCTACAGC_EcoRI_P1.Bottom       | 5'-Phos-AATTCTGATAGCGATCGGAAGAGCGTCTCGTGTAGGGAAAGAGTGT-3' |
| GTACATCA_EcoRI_P1.Top          | 5'-ACACTCTTCTCCCTACACGAGCAGCTCTCCCGATCTGTACATCA-3'   |
| GTACATCA_EcoRI_P1.Bottom       | 5'-Phos-AATTCTGATAGCGATCGGAAGAGCGTCTCGTGTAGGGAAAGAGTGT-3' |
| GTGTACTG_EcoRI_P1.Top          | 5'-ACACTCTTCTCCCTACACGAGCAGCTCTCCCGATCTGTACTG-3'    |
| GTGTACTG_EcoRI_P1.Bottom       | 5'-Phos-AATTCTGATAGCGATCGGAAGAGCGTCTCGTGTAGGGAAAGAGTGT-3' |
| Oligo ID | Sphl Adapter P2.1_Oligo sequence | Sphl Adapter P2.2_Oligo sequence |
|----------|---------------------------------|---------------------------------|
| TACGATAT_EcoRI_P1.Top | 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTACGATAT | |
| TACGATAT_EcoRI_P1.Bottom | 5'-Phos-AATTATATCGTAAGATCGGAAGAGCGGTGCTGTAAGGAAAAGAGTGT-3' | |
| TCAGCATC_EcoRI_P1.Top | 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAGCATC | |
| TCAGCATC_EcoRI_P1.Bottom | 5'-Phos-AATTGATGCTGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3' | |
| TCGAGTGA_EcoRI_P1.Top | 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAGGAGTGTA | |
| TCGAGTGA_EcoRI_P1.Bottom | 5'-Phos-AATTTCACCTCGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3' | |
| TCTCTCGA_EcoRI_P1.Top | 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCTCTCGA | |
| TCTCTCGA_EcoRI_P1.Bottom | 5'-Phos-AATTTCAGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3' | |
| TGCACTAC_EcoRI_P1.Top | 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGCACTAC | |
| TGCACTAC_EcoRI_P1.Bottom | 5'-Phos-AATTGTAAGTGCAGAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3' | |
| TGGACTG_EcoRI_P1.Top | 5'-ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACTG | |
| TGGACTG_EcoRI_P1.Bottom | 5'-Phos-AATTCAGTCACAAGATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3' | |
| Sphl-P2_Top | 5'-Phos-AGATCGGAAGAGCGGAAGACACCA-3' | |
| Sphl-P2_Bottom | 5'-GTGACTGAGTTTCAGACGTTGCTCTTCGATCTCATG-3' | |
Oligo ID
P1 PCR primer
Oligo sequence
5'-AATGATACGGCGACCACCGAGATCTACACTTTCCCTACACGACG-3'

Oligo ID
P2 PCR primer
Oligo sequence
5'-CAAGCAGAAGACGGCATACGAGATAACGTGATGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_AACGTGAT
5'-CAAGCAGAAGACGGCATACGAGATACACTTTCCCTACACGACG-3'

Oligo ID
ddRAD_PCR2_Idx_CGCTGATC
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_ACAAGCTA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_CTGTAGCC
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_ACGCTCGA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_ ACGTATCA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_ ACTATGCA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_AGAGTCAA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_AGATCGCA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_AGCAGGAA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_AGTCACTA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_ATCCTGTA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_ATTGAGGA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_CACTTCGA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_CAGCGTTA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_CATACCAA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_CCAGTTCA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_GTGTTCTA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_GCTCGGTA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_GGAGAACA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_GTGTTCTA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_TAGGATGA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_TGGCTTCA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_TTCACGCA
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'

Oligo ID
ddRAD_PCR2_Idx_GCCAAGAC
5'-CAAGCAGAAGACGGCATACGAGATACGTATCAGTGACTGGAGTTCAGACGTGTGC-3'
| Population by species | Number of SNPs | $F_{IS}$  |
|-----------------------|----------------|---------|
| *G. rivale*            |                |         |
| Ben Lawers            | 2531           | -0.0082 |
| Coire Garblach        | 2605           | 0.0384  |
| Ben Lui               | 2664           | 0.0423  |
| *G. urbanum*          |                |         |
| Priory Wood           | 551            | 0.6216  |
| Mill Wood             | 531            | -0.0697 |
| Burgh Wood            | 542            | 0.5913  |
| Hoades Wood           | 451            | 0.5100  |
| Punnetts Town         | 495            | 0.4636  |
| Frith Wood            | 507            | 0.6134  |
| Stanford Bridge       | 535            | 0.6112  |
| Combe Wood            | 536            | 0.6110  |
| Copperhurst           | 509            | 0.5874  |
| Selwyn Wood           | 166            | -0.4759 |

Table S4. $F_{IS}$ estimates for ‘allopatric’ UK *G. rivale* and *G. urbanum* populations using data not filtered for paralogs.
| Sample     | Heterozygosity | No. SNPs private to *G. rivale* | No. SNPs private to *G. urbanum* | No. SNPs shared by both species | No. alternately fixed SNPs | No. DNA blocks |
|------------|----------------|---------------------------------|-----------------------------------|---------------------------------|----------------------------|----------------|
| Ben Lui 1  | 250            | 88                              | 0                                 | 1184                            | 668                        |                |
| Ben Lui 4  | 299            | 86                              | 1                                 | 1170                            | 672                        |                |
| Ben Lawers 5 | 179           | 88                              | 0                                 | 1189                            | 649                        |                |

Table S5: Polymorphism counts and numbers of DNA blocks for coalescent analysis for three individual *G. rivale* samples from UK allopatric populations.
| Model | Model 1 | Model 2 | Model 3 |
|-------|---------|---------|---------|
| \( N_{\text{anc}} = N_{\text{rivb}} \) | 1.03 * 10^5 | 1.01 * 10^5 | 6.42 * 10^4 |
| \( N_{\text{urb}} \) | 3.46 * 10^4 | 3.23 * 10^4 | 3.29 * 10^4 |
| \( t \) (years) | 2.36 * 10^6 | 2.41 * 10^6 | 3.03 * 10^6 |
| \( M \) | 0.02 (0.00 – 0.18) | | 0.08 (0.06 – 0.11) |

**Ben Lui 4**

| \( N_{\text{anc}} = N_{\text{rivb}} \) | 1.30 * 10^5 | 1.24 * 10^5 | 1.11 * 10^5 |
| \( N_{\text{urb}} \) | 3.77 * 10^4 | 3.09 * 10^4 | 3.38 * 10^4 |
| \( t \) (years) | 2.22 * 10^6 (2.01 * 10^6 – 2.43 * 10^6) | 2.35 * 10^6 | 2.48 * 10^6 |
| \( M \) (mean, 95% CI’s) | 0.04 (0.005 – 0.191) | | 0.04 (0.007 – 0.101) |

**Ben Lawers 5**

| \( N_{\text{anc}} = N_{\text{rivb}} \) | 7.41 * 10^4 | 7.34 * 10^4 | 5.11 * 10^4 |
| \( N_{\text{urb}} \) | 3.54 * 10^4 | 3.40 * 10^4 | 3.38 * 10^4 |
| \( t \) (years) | 2.60 * 10^6 | 2.64 * 10^6 | 3.10 * 10^6 |
| \( M \) (mean, 95% CI’s) | 0.01 (0.00 – 0.103) | | 0.02 (0.002 – 0.052) |

| Parameter estimates under models analysed for sample pairs involving the samples: Ben Lui 1, Ben Lui 4, and Ben Lawers 5. 95% Confidence intervals provided in parentheses. See Figure S4 for plots of 95% CI’s. |
Figure S1: Bioinformatics pipeline for analysis of Geum ddRAD sequences.
Figure S2.
Figure S3

Histogram of Coverage per Scaffold

Coverage per scaffold

Frequency

0  5000
1000 4000
2000
3000
4000
5000

0  10  20  30  40  50
Figure S4: Log-Likelihood plots for estimates of M for three pairs of *Geum* samples, each involving a different *G. rivale* individual (Ben Lui 1, Ben Lui 4, Ben Lawers 5). Blue and green lines represent log-likelihoods of M for models IM*ru* and IM*ur*, respectively. 95% CI’s are bounded where these lines intersect the dark horizontal line.
Figure S5. Number of species-specific SNPs scored per individual in a sample of 97 Geum plants from 39 populations in Berwickshire.
Figure S6. Genomic composition estimates for Berwickshire samples derived from paralog-filtered species-diagnostic SNPs (y-axis in both panels) compared to analogous measures in two alternate analyses: (A) species-diagnostic SNPs from data unfiltered for paralogs, and (B) Admixture proportion (Q value) calculated by fastSTRUCTURE analysis that required loci to be present in all individuals (188 SNPs).
Figure S7. Number of putatively introgressed loci per individual in individuals classified as *G. rivale* (left panel) and *G. urbanum* (right panel) sampled from 39 populations in Berwickshire.