# The *Aquilegia* genome: adaptive radiation and an extraordinarily polymorphic chromosome with a unique history

**Danièle Filiault[1], Evangeline S. Ballerini[2], Terezie Mandáková[3], Gökçe Aköz[1], Nathan Derieg[2], Jeremy Schmutz[4,5], Jerry Jenkins[4,5], Jane Grimwood[4,5], Shengqiang Shu[4], Richard D. Hayes[4], Uffe Hellsten[4], Kerrie Barry[4], Juying Yan[4], Sirma Mihaltcheva[4], Miroslava Karafiátová[6], Viktoria Nizhynska[1], Martin A. Lysak[3], Scott A. Hodges[2,*], and Magnus Nordborg[1,*]**

[1]Gregor Mendel Institute, Austrian Academy of Sciences, Vienna Biocenter, Vienna, Austria

[2]Department of Ecology, Evolution, and Marine Biology, University of California Santa Barbara, Santa Barbara, USA

[3]CEITEC - Central-European Institute of Technology, Masaryk University, Brno, Czech Republic

[4]Department of Energy Joint Genome Institute, Walnut Creek, California, USA

[5]HudsonAlpha Institute of Biotechnology, Huntsville, Alabama, USA

[6]Institute of Experimental Botany, Centre of the Region Haná for Biotechnological and Agricultural Research, Olomouc, Czech Republic

[*]scott.hodges@lifesci.ucsb.edu, magnus.nordborg@gmi.oeaw.ac.at

## ABSTRACT

The columbine genus *Aquilegia* is a classic example of an adaptive radiation, involving a wide variety of pollinators and habitats. Here we present the genome assembly of *A. coerulea* 'Goldsmith', complemented by high-coverage sequencing data from 10 wild species covering the world-wide distribution. Our analysis reveals extensive allele sharing among species, and sheds light on the complex process of radiation. Our analysis also led to the remarkable discovery that the evolutionary history of an entire chromosome differed from that of the rest of the genome – a phenomenon which we do not fully understand, but which highlights the need to consider chromosomes in an evolutionary context.

## Introduction

Understanding adaptive radiation is a longstanding goal of evolutionary biology[1]. As a classic example of adaptive radiation, the *Aquilegia* genus has outstanding potential as a subject of such evolutionary studies[2–4]. The genus is made up of about 70 species distributed across Asia, North America, and Europe[5] (**Fig. 1**). Distributions of many *Aquilegia* species overlap or adjoin one another, sometimes forming notable hybrid zones[6,7]. Additionally, species tend to be widely interfertile, especially within geographic regions[8].

Phylogenetic studies have defined two concurrent, yet contrasting, adaptive radiations in *Aquilegia*[9,10]. From a common ancestor in Asia, one radiation occurred in North America via Northeastern Asian precursors, while a separate Eurasian radiation took place in central and western Asia and Europe. While adaptation to different habitats is thought to be a common force driving both radiations, shifts in primary pollinators also play a substantial role in North America[9,11]. Previous phylogenetic studies have frequently revealed polytomies[6,9–12], suggesting that many *Aquilegia* species are very closely related.

Genomic data are beginning to uncover the extent to which interspecific variant sharing reflects a lack of strictly bifurcating species relationships, particularly in the case of adaptive radiation. Discordance between gene and species trees has been widely observed[13] (and references 15,34-44 therein)[14,15], and while disagreement at the level of individual genes is expected under standard population genetics coalescent models[16] (also known as "incomplete lineage sorting"[17]), there is increased evidence for systematic discrepancies that can only be explained by some form of gene flow[13–15,18]. The importance of admixture as a source of adaptive genetic variation has also become more evident[19–21]. Hence, rather than being a problem to overcome in phylogenetic analysis, non-bifurcating species relationships could actually describe evolutionary processes that are fundamental to understanding speciation itself. Here we generate an *Aquilegia* reference genome based on the horticultural cultivar *Aquilegia coerulea* 'Goldsmith' and perform resequencing and population genetic analysis of 10 additional individuals representing North American, Asian, and European species, focusing in particular on the relationship between species.

# Results

## Genome assembly

We sequenced an inbred horticultural cultivar (*A. coerulea* 'Goldsmith') using a whole genome shotgun sequencing strategy (See **Materials and Methods** for more details). With the aid of genetic maps, we assembled sequences into a 291.7 Mb reference genome consisting of 7 chromosomes (282.6 Mbp) and an additional 1,027 scaffolds (9.13 Mbp). Genes were annotated using RNAseq data of a variety of tissues and species (**Supplementary Table 1**), EST data sets[22], and protein homology support, yielding 30,023 loci and 13,527 alternate transcripts. The *A. coerulea* v3.1 genome release is available on Phytozome (https://phytozome.jgi.doe.gov/pz/portal.html).

## Polymorphism and divergence

We deeply resequenced one individual from each of ten *Aquilegia* species (**Fig. 1**). Sequences were aligned to the *A. coerulea* v3.1 reference using bwa mem[23,24] and variants were called using GATK Haplotype Caller[25]. Genomic positions were conservatively filtered to identify the portion of the genome in which variants could be reliably called across all ten species (see *Materials and Methods* for alignment, SNP calling, and genome filtration details). The resulting callable portion of the genome was heavily biased towards genes and included 57% of annotated coding regions (48% of gene models), but only 21% of the reference genome as a whole.

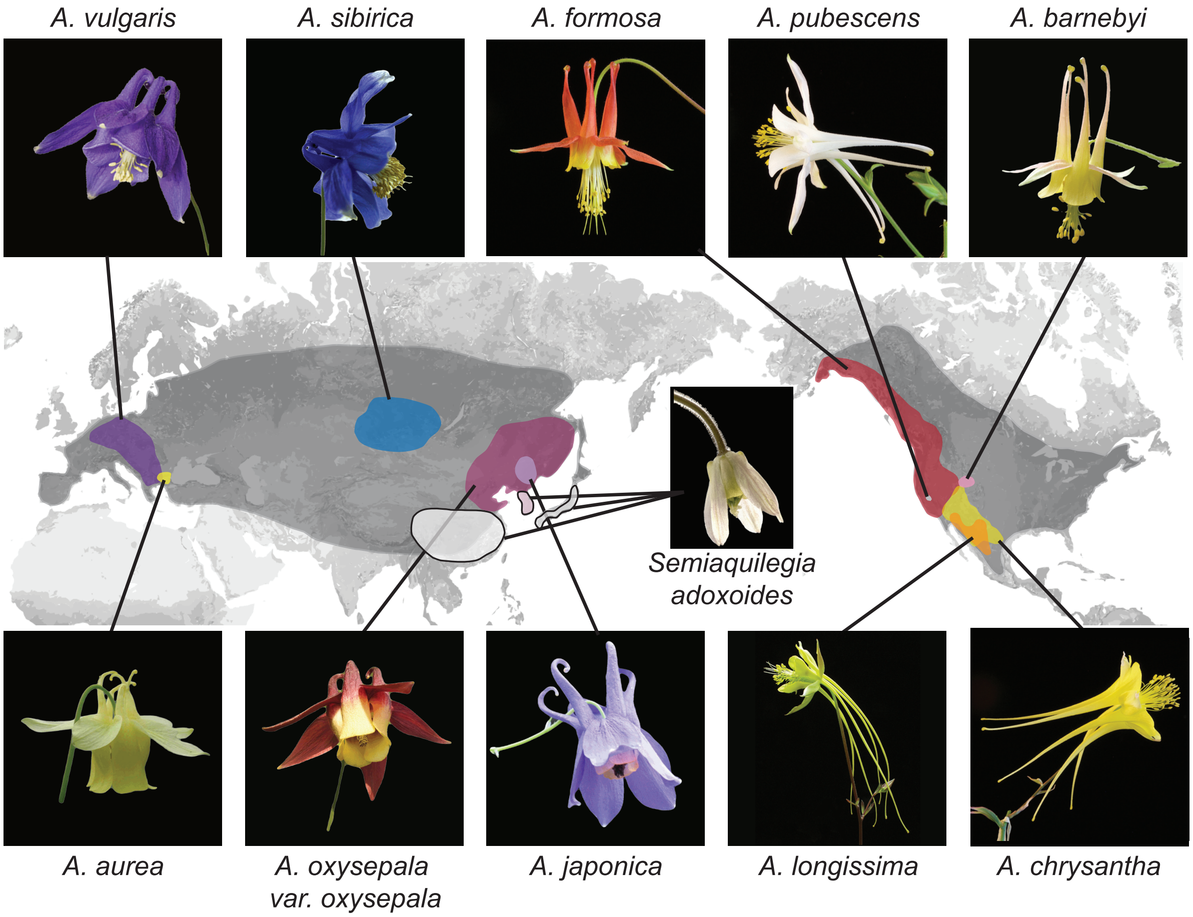


**Figure 1. Distribution of *Aquilegia* species.** There are ~70 species in the genus *Aquilegia*, broadly distributed across temperate regions of the Northern Hemisphere (grey). The 10 *Aquilegia* species sequenced here were chosen as representatives spanning this geographic distribution as well as the diversity in ecological habitat and pollinator-influenced floral morphology of the genus. *Semiaquilegia adoxoides*, generally thought to be the sister taxon to *Aquilegia*[10], was also sequenced. A representative photo of each species is shown and is linked to its approximate distribution.

Using these callable sites, we calculated nucleotide diversity as the percentage of pairwise sequence differences in each individual. Assuming random mating, this metric reflects both individual sample heterozygosity and nucleotide diversity in the species as a whole. Of the ten individuals, most had a nucleotide diversity of 0.2-0.35% (**Fig. 2a**), similar to previous estimates of nucleotide diversity in *Aquilegia*[26], yet lower than that of a typical outcrossing species[27]. While likely partially attributable to enrichment for highly conserved genomic regions with our stringent filtration, these lower-than-expected heterozygosity levels could also reflect inbreeding. Additionally, four individuals in our panel had extended stretches of very low levels of heterozygosity (nucleotide diversity < 0.1%) consistent with recent inbreeding (**Supplementary Fig. 1**). Selfing does appear to be common in *Aquilegia*, but estimates of inbreeding in adults are generally low, suggesting substantial inbreeding depression[28–30].

We next considered nucleotide diversity between individuals as a measure of species divergence. Divergence within a region (0.38-0.86%) was often only slightly higher than within-species diversity, implying extensive variant sharing, while divergence between regions was markedly higher (0.81-0.97%; **Fig.2a**). $F_{ST}$ between regions (0.245-0.271) was similar to that between common *Arabidopsis* species[13], yet lower than between most vervet species pairs[14], and higher than between cichlid groups in Malawi[31] or human ethnic groups[32]. The topology of trees constructed with concatenated genome data (neighbor joining (**Fig. 2a**), RAxML (**Supplementary Fig. 2**)) were in broad agreement with previous *Aquilegia* phylogenies[9–12, 33], with one exception: while *A. oxysepala* is basal in our analysis, it had been placed within the large Eurasian clade with moderate to strong support in previous studies[9, 10].

Surprisingly, levels of polymorphism were generally strikingly higher on chromosome four (**Fig. 2b**. Exceptions were apparently due to inbreeding, especially in the case of the *A. aurea* individual, which appears to be almost completely homozygous (**Fig. 2a** and **Supplementary Fig. 1**). The increased polymorphism on chromosome four is only partly reflected in increased divergence to an outgroup species (*Semiaquilegia adoxoides*), suggesting that it represents deeper coalescence times rather than simply a higher mutation rate (mean ratio chromosome four/genome: polymorphism=2.258, divergence=1.201, **Supplementary Table 2**).

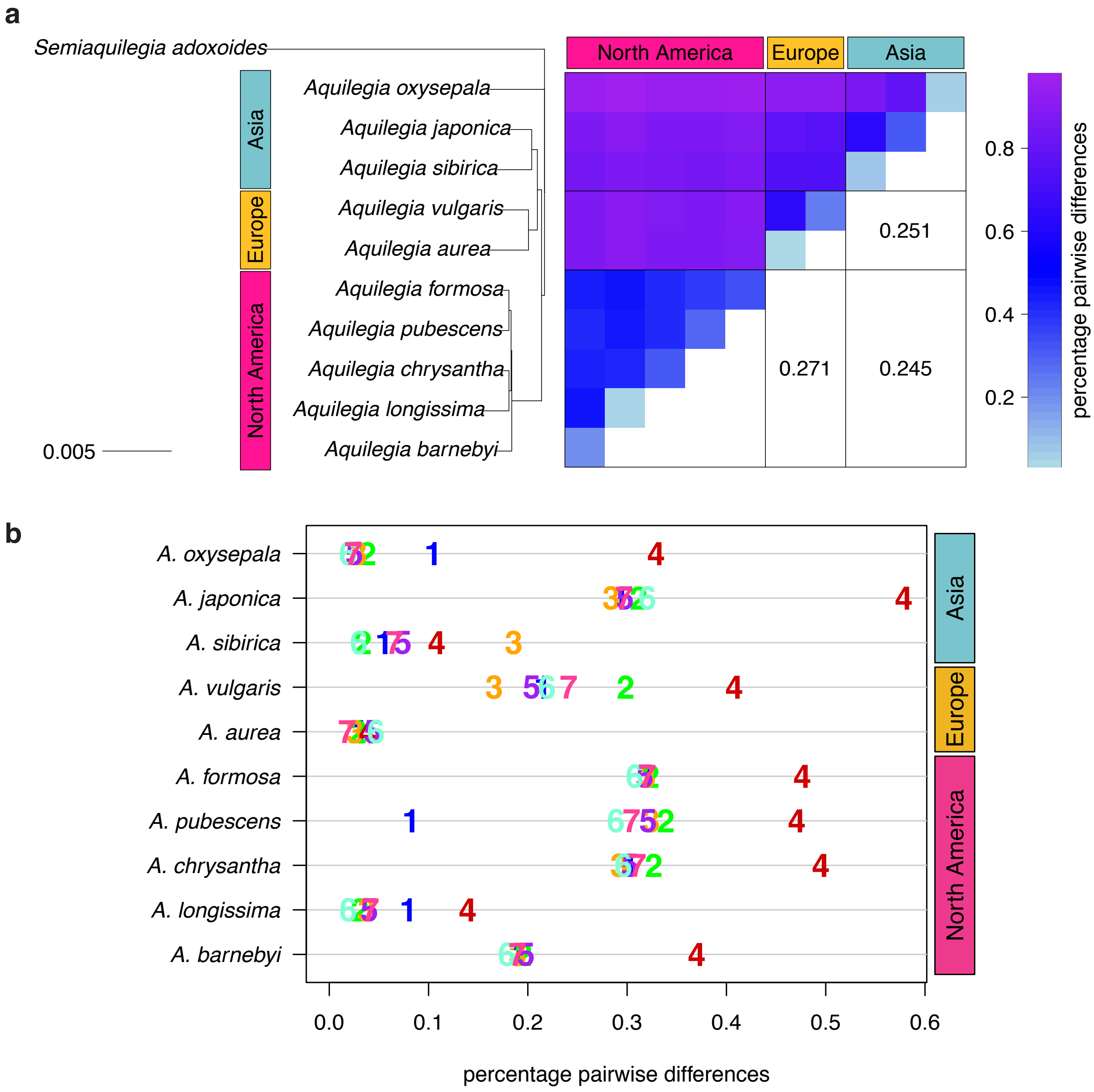


**Figure 2. Polymorphism and divergence in *Aquilegia*.** (a) The percentage of pairwise differences within each species (estimated from individual heterozygosity) and between species (divergence). $F_{ST}$ values between geographic regions are given on the lower half of the pairwise differences heatmap. Both heatmap axes are ordered according to the neighbor joining tree to the left. This tree was constructed from a concatenated data set of reliably-called genomic positions. (b) Polymorphism within each sample by chromosome. Per chromosome values are indicated by the chromosome number.

## Discordance between gene and species trees

To assess discordance between gene and species trees, we constructed a cloudogram of trees drawn from 100kb windows across the genome (**Fig. 3a**). Fewer than 1% of these window-based trees were topologically identical to the species tree. North American species were consistently separated from all others (96% of window trees) and European species were also clearly delineated (67% of window trees). However, three bifurcations delineating Asian species were much less common: the *A. japonica* and *A. sibirica* sister relationship (45% of window trees), a basal placement of *A. oxysepala* (30% of window trees), and the split demarcating the Eurasian radiation (31% of window trees). These results demonstrate a marked discordance of gene and species trees throughout both *Aquilegia* radiations.

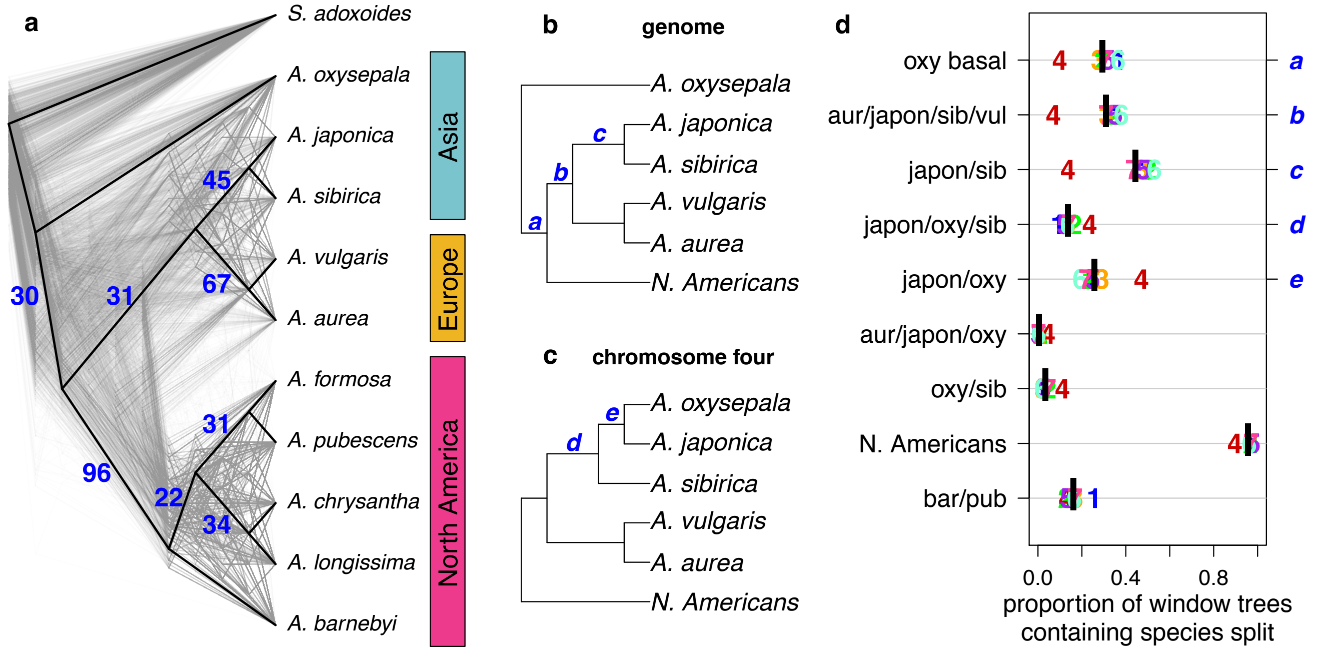


**Figure 3. Discordance between gene and species trees.** (a) Cloudogram of neighbor joining (NJ) trees constructed in 100kb windows across the genome. The topology of each window-based tree is co-plotted in grey and the whole genome NJ tree shown in **Fig. 2a** is superimposed in black. Blue numbers indicate the percentage of window trees that contain each of the subtress observed in the whole genome tree. (b) Genome NJ tree topology. Blue letters a-c on the tree denote subtrees a-c in panel (d). (c) Chromosome four NJ tree topology. Blue letters d and e on the tree denote subtrees d and e in panel (d). (d) Prevalence of each subtree that varied significantly by chromosome. Genomic (black bar) and per chromosome (chromosome number) values are given.

The gene tree analysis also reflected the unique evolutionary history of chromosome 4. Of 217 unique subtrees observed in gene trees, nine varied significantly in frequency between chromosomes (chi-square test $p$-value $< 0.05$ after Bonferroni correction; **Fig. 3b** and **Supplementary Tables 3 and 4**). Trees describing a sister species relationship between *A. pubescens* and *A. barnebyi* were more common on chromosome one, but chromosome four stood out with respect to eight other relationships, most of them related to *A. oxysepala* (**Fig. 3d**). Although *A. oxysepala* was basal in our genome tree, the topology of the chromosome four tree was consistent with previously-published phylogenies in that it placed *A. oxysepala* within the Eurasian clade[9, 10](**Supplementary Fig. 2**). Subtree prevalences were in accordance with this topological variation (**Fig. 3b-d**).

The subtree delineating all North American species was also less frequent on chromosome four, indicating that the history of the chromosome is discordant in both radiations. We detected no patterns in the prevalence of any chromosome-discordant subtree that would suggest structural variation or a large introgression (**Supplementary Fig. 3**).

## Polymorphism sharing across the genus

We next polarized variants against an outgroup species (*S. adoxoides*) to explore the prevalence and depth of polymorphism sharing. Private derived variants accounted for only 21-25% of polymorphic sites in North American species and 36-47% of variants in Eurasian species (**Fig. 4a**). The depth of polymorphism sharing reflected the two geographically-distinct radiations. North American species shared 34-38% of their derived variants within North America, while variants in European and Asian species were commonly shared across two geographic regions (18-22% of polymorphisms, predominantly shared between Europe and Asia; **Fig. 4b and c**; **Supplementary Table 5**). Strikingly, a large percentage of derived variants occurred in all three geographic regions (22-32% of polymorphisms, **Fig. 4d**), demonstrating that polymorphism sharing in *Aquilegia* is extensive and deep.

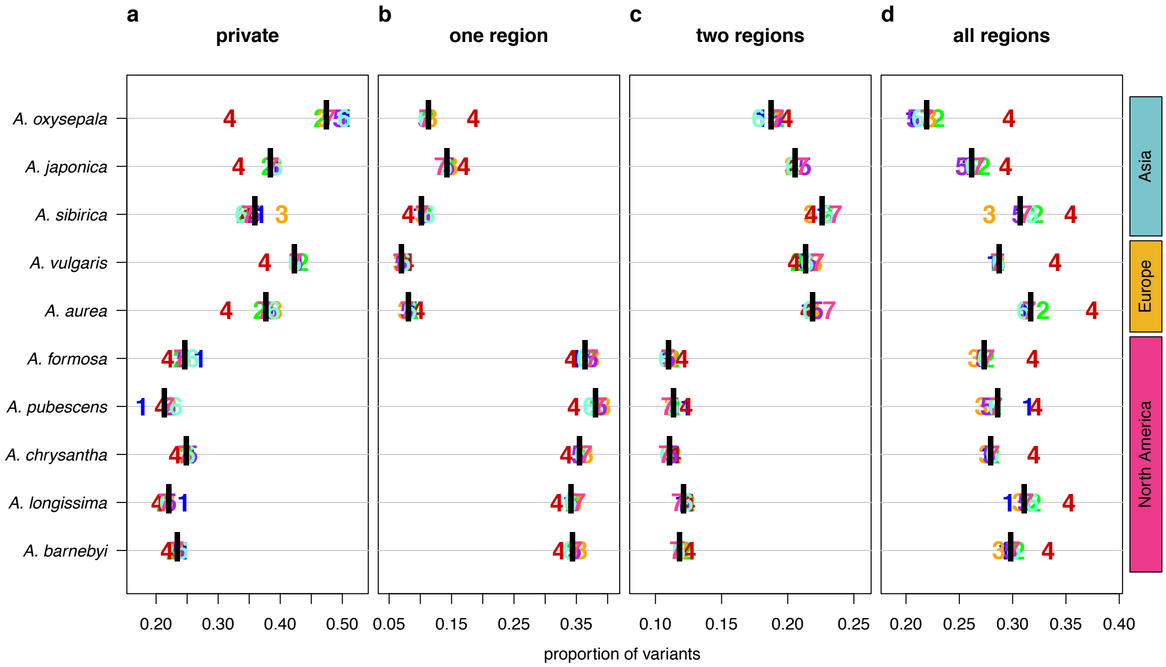


**Figure 4. Sharing patterns of derived polymorphisms.** Proportion of derived variants (a) private to an individual species, (b) shared within the geography of origin, (c) shared across two geographic regions, and (d) shared across all three geographic regions. Genomic (black bar) and chromosome (chromosome number) values, for all 10 species.

In all species examined, the proportion of deeply shared variants was higher on chromosome four (**Fig. 4d**), largely due to a reduction in private variants, although sharing at other depths was also reduced in some species. Variant sharing on chromosome four within Asia was higher in both *A. oxysepala* and *A. japonica* (Fig. 4b), primarily reflecting higher variant sharing between these species (**Fig. 6a**).

## Evidence of gene flow

Consider three species, H1, H2, and H3. If H1 and H2 are sister species relative to H3, then, in the absence of gene flow, H3 must be equally related to H1 and H2. The D statistic[18] tests this hypothesis by comparing the number of derived variants shared between H3, and H1 and H2, respectively. A non-zero D statistic reflects an asymmetric pattern of allele sharing, implying gene flow between H3 and one of the two sister species, i.e., that speciation was not accompanied by complete reproductive isolation. If *Aquilegia* diversification occurred via a series of bifurcating species splits characterized by reproductive isolation, bifurcations in the species tree should represent combinations of basal and derived species with symmetric allele sharing patterns (D=0). Given the high discordance of gene and species trees at the individual species level, we focused on testing a simplified tree topology based on the three groups whose bifurcation order seemed clear: (1) North American species, (2) European species, and (3) Asian species not including *A. oxysepala*. *S. adoxoides* was used to determine the ancestral state of alleles in all tests.

We first tested each North American species as H3 against all combinations of European and Asian (without *A. oxysepala*) species as H1 and H2 (**Fig. 5a-c**). As predicted, the North American split was closest to resembling speciation with strict reproductive isolation, with little asymmetry in allele sharing between North American and Asian species and low, but significant, asymmetry between North American and European species (**Fig. 5b**). Next, we considered allele sharing between European and Asian (without *A. oxysepala*) species (**Fig. 5 d and e**). Here we found non-zero D-statistics for all species combinations.

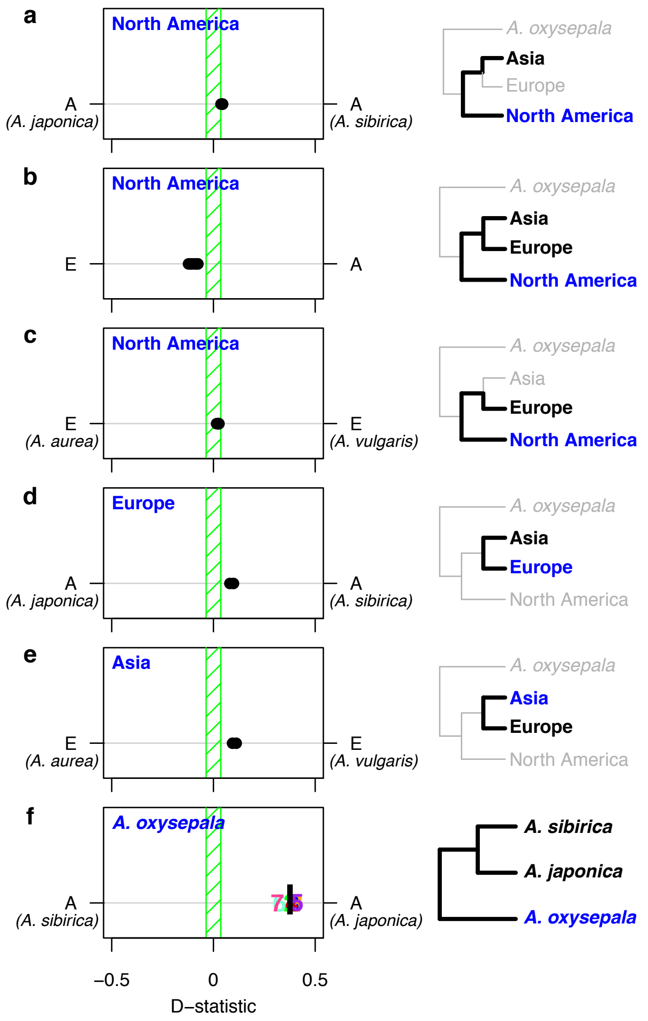
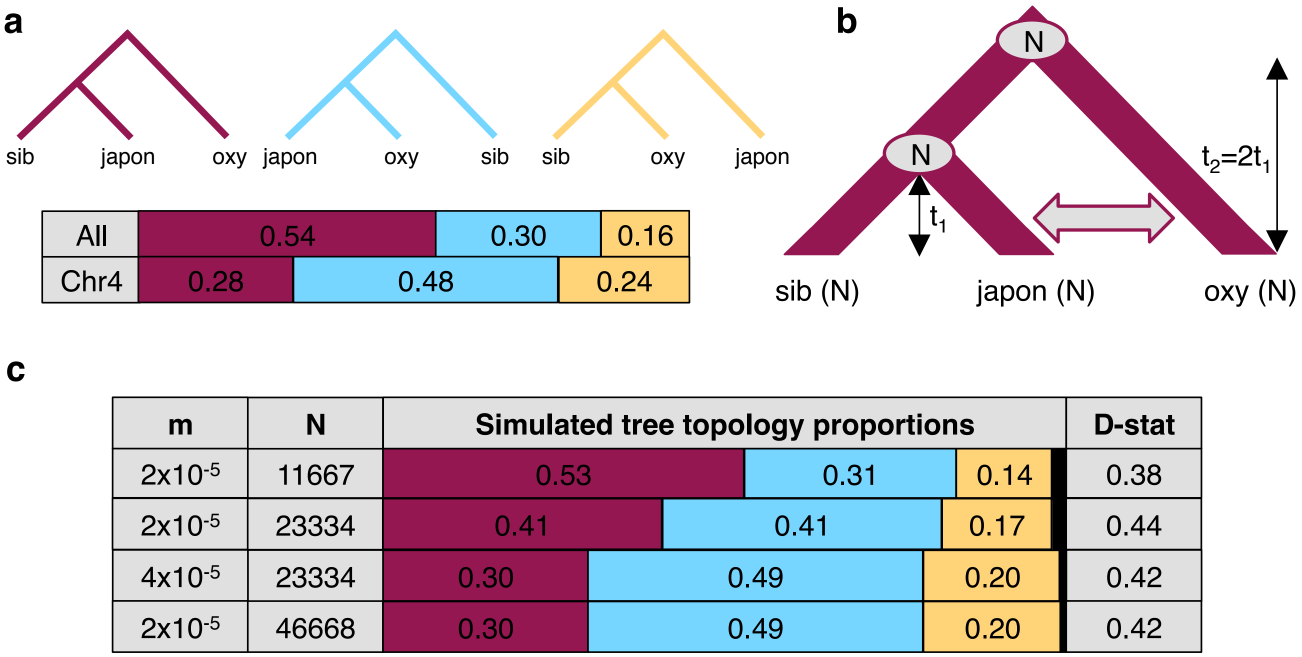


**Figure 5. D statistics demonstrate gene flow during *Aquilegia* speciation.** D statistics for tests with (a-c) all North American species, (d) both European species, (e) Asian species other than *A. oxysepala*, and (f) *A. oxysepala* as H3 species. All tests use *S. adoxoides* as the outgroup. D statistics outside the green shaded areas are significantly different from zero. In (a-e), each individual dot represents the D statistic for a test done with a unique species combination. In (f), D statistics are presented by chromosome (chromosome number) or by the genome-wide value (black bar). In all panels, E=European and A=Asian without *A. oxysepala*. In some cases, individual species names are given when the geographical region designation consists of a single species. Right hand panels are a graphical representation of the D statistic tests in the corresponding left hand panels. Trees are a simplified version of the genome tree topology (**Fig. 2b**), in which the bold subtree(s) represent the bifurcation considered in each set of tests. H3 species are noted in blue while the H1 and H2 species are specified in black.

Interestingly, the patterns of asymmetry between these two regions were reticulate: Asian species shared more variants with the European *A. vulgaris* while European species shared more derived alleles with the Asian *A. sibirica*. D statistics therefore demonstrate widespread asymmetry in variant sharing between *Aquilegia* species, suggesting that speciation processes throughout the genus were not characterized by strict reproductive isolation.

Although non-zero D statistics are usually interpreted as being due to gene flow in the form of admixture between species, they can also result from gene flow between incipient species. Either way, speciation precedes reproductive isolation. The possibility that different levels of purifying selection in H1 or H2 explain the observed D statistics can probably be ruled out, since D statistics do not differ when calculated with only fourfold degenerate sites ($p$-value < $2.2 \times 10^{-16}$, adjusted $R^2$=0.9942, **Supplementary Table 6**). Non-zero D statistics could also indicate that the bifurcation order tested was incorrect, but even tests based on alternative tree topologies resulted in few D statistics that equal zero (**Supplementary Table 6**). Therefore, the non-zero D statistics observed in *Aquilegia* most likely reflect a pattern of reticulate evolution throughout the genus.

Since variant sharing between *A. oxysepala* and *A. japonica* was higher on chromosome four (**Fig. 6a**), and hybridization between these species has been reported[7] we wondered whether gene flow could explain the discordant placement of *A. oxysepala* between chromosome four and genome trees (**Fig. 3b and c**). Indeed, when the genome tree was taken as the bifurcation order, D statistics were elevated between these species (**Fig. 5f**). A relatively simple model allowing for bidirectional gene flow between *A. oxysepala* and *A. japonica* (**Fig. 6b**) demonstrated that doubling the population size (N) to reflect chromosome four's polymorphism level (i.e. halving the coalescence rate) could indeed shift tree topology proportions (**Fig. 6c**, row 2). However, recreating the observed allele sharing ratios on chromosome four (**Fig. 6a**) required some combination of increased migration (m) and/or N (**Fig. 6c**, rows 3-4). It is plausible that gene flow might differentially affect chromosome four, and we will return to this topic in the next section. Although the similarity of the D statistic across chromosomes (**Fig. 5f**) might seem inconsistent with increased migration on chromosome four, the D statistic reaches a plateau in our simulations such that many different combinations of m and N produce similar D values (**Fig 6c** and **Supplementary Fig. 4**). This underscores the idea that D statistics are a general indicator of gene flow, but not necessarily a quantitative measure of its magnitude. Therefore, an increase in migration rate and deeper coalescence can explain the tree topology of chromosome four, a result that might explain inconsistencies in *A. oxysepala* placement in previous phylogenetic studies[9,10].



| m | N | Simulated tree topology proportions | | | D-stat |
|---|---|---|---|---|---|
| $2x10^{-5}$ | 11667 | 0.53 | 0.31 | 0.14 | 0.38 |
| $2x10^{-5}$ | 23334 | 0.41 | 0.41 | 0.17 | 0.44 |
| $4x10^{-5}$ | 23334 | 0.30 | 0.49 | 0.20 | 0.42 |
| $2x10^{-5}$ | 46668 | 0.30 | 0.49 | 0.20 | 0.42 |

**Figure 6. The effect of differences in coalescence time and gene flow on tree topologies.** (a) The observed proportion of informative derived variants supporting each possible Asian tree topology genome-wide and on chromosome four. Species considered include *A. oxysepala* (oxy), *A. japonica* (japon), and *A. sibirica* (sib). (b) The coalescent model with bidirectional gene flow in which *A. oxysepala* diverges first at time t2 ("species tree"), but later hybridizes with *A. japonica* between t=0 and t1 at a rate determined by per-generation migration rate, m. The population size (N) remains constant at all times. (c) The proportion of each tree topology and estimated D statistic for simulations using five combinations of m and N values (t1=1 in units of N generations). The combination presented in the first row (m=2x10$^{-5}$ and N=11667) generates tree topology proportions that match observed allele sharing proportions genomewide. Simulations with increased m and/or N (rows 3-4) result in proportions which more closely resemble those observed for chromosome four. Colors in proportion plots refer to tree topologies in (a), with black bars representing the residual probability of seeing no coalescence event. While this simulation assumes symmetric gene flow, similar results were seen for models incorporating both unidirectional and asymmetric gene flow (**Supplementary Fig. 4** and **Supplementary Table 7**).

## The pattern of polymorphism on chromosome four

In most of the sequenced *Aquilegia* species, the level of polymorphism on chromosome four is twice as high as in the rest of the genome (**Fig. 2b**). This unique pattern could be: 1) an artifact of biases in polymorphism detection between chromosomes, 2) the result of a higher neutral mutation rate on chromosome four, or 3) the result of deeper coalescence times on chromosome 4 (allowing more time for polymorphism to accumulate).

While it is impossible to completely rule out phenomena such as cryptic copy number variants, for the pattern to be entirely attributable to artefacts would require that half of the polymorphism on chromosome 4 be spurious. This scenario is extremely unlikely given our extensive quality control and filtering (see **Supplementary Table 8**). Similarly, the pattern cannot wholly be explained by a higher neutral mutation rate. If this were the case, both divergence and polymorphism would be elevated to the same extent on chromosome four[34]. As noted above, this not the case (**Supplementary Table 2**). Thus the higher level of polymorphism on chromosome four must to some extent reflect differences in coalescence time, which can only be due to selection.

Although it is clear that selection can have a dramatic effect on the history of a single locus, the chromosome-wide pattern we observe (**Supplementary Fig. 1**) is difficult to explain. It cannot be due to selection on a limited number of linked loci, as we saw no pattern of recombination on chromosome four that would suggest the presence of a supergene or inversion (*A. formosa* x *A. pubescens* F2 mapping population; **Supplementary Fig. 5**). Selection must thus be acting on a very large number of loci across the chromosome. Balancing selection is known to elevate polymorphism, and chromosome four is enriched for defense genes (**Supplementary Table 9**), some of which have been shown to be under balancing selection[35,36]. However, while significant, this enrichment involves a relatively small number of genes and is therefore unlikely to completely explain the polymorphism pattern[37].

Another potential explanation is so-called background selection[38]. Several characteristics of chromosome four suggest that it could experience less purifying selection than the rest of the genome. Gene density is markedly lower (**Supplementary Table 10**), it harbors a higher proportion of repetitive sites (**Supplementary Table 10**), and is enriched for many transposon families such as Copia and Gypsy elements, among others (**Supplementary Table 11**). Additionally, a higher proportion of genes on chromosome four were either not expressed or expressed at a low level (**Supplementary Fig. 6**). Gene models on the chromosome were also more likely to contain variants that could disrupt protein function (**Supplementary Table 10**). Taken together, these observations suggest less purifying selection on chromosome 4. However, there is no evidence of this leading to reduced background selection, as we observed no significant correlation between gene density and polymorphism on any chromosome (**Supplementary Table 12**). Reduced purifying selection could also explain the prediction of higher gene flow between *A. oxysepala* and *A. japonica* on chromosome four (**Fig. 6**); the chromosome would be more permeable to gene flow if loci involved in the adaptive radiation were preferentially located on other chromosomes.

While selection during the *Aquilegia* radiation contributes to the pattern of polymorphism on chromosome four, the pattern itself predates the radiation. Divergence between *Aquilegia* and *Semiaquilegia* is higher on chromosome four (2.77%

on chromosome four, 2.48% genome-wide, **Supplementary Table 13**), as is heterozygosity within *Semiaquilegia* (0.16% chromosome four, 0.08% genome-wide, **Supplementary Table 13**). This suggests that the variant evolutionary history of chromosome four originated before the *Aquilegia*/*Semiaquilegia* split.

## The 35S and 5S rDNA loci are uniquely localized to chromosome four

The observation that one *Aquilegia* chromosome is different from the others is not novel; previous cytological work described a single nucleolar chromosome that appeared to be highly heterochromatic[39]. Using fluorescence in situ hybridization (FISH) with rDNA and chromosome four-specific bulked oligo probes[40], we confirmed that both the 35S and 5S rDNA loci were localized uniquely to chromosome four in two *Aquilegia* species and *S. adoxoides* (**Fig. 7**). The chromosome contained a single large 35S repeat cluster proximal to the centromeric region in all three species. Interestingly, the 35S locus in *A. formosa* was larger than that of the other two species and formed variable bubbles and fold-backs on extended pachytene chromosomes similar to structures previously observed in *Aquilegia* hybrids[39] (**Fig. 7**, last panels). The 5S rDNA locus was also proximal to the centromere on chromosome four, although slight differences in the number and position of the 5S repeats between species highlight the dynamic nature of this gene cluster. However, no chromosome appeared to be more heterochromatic than others in our analyses (**Fig. 7**); FISH with 5-methylcytosine antibody showed no evidence for hypermethylation on chromosome four (**Supplementary Fig. 7**) and GC content was similar for all chromosomes (**Supplementary Table 14**). However, similarities in chromosome four organization across all three species reinforce the idea that the exceptionality of this chromosome predated the *Aquilegia*/*Semiaquilegia* split and raise the possibility that rDNA clusters could have played a role in the variant evolutionary history of chromosome four.

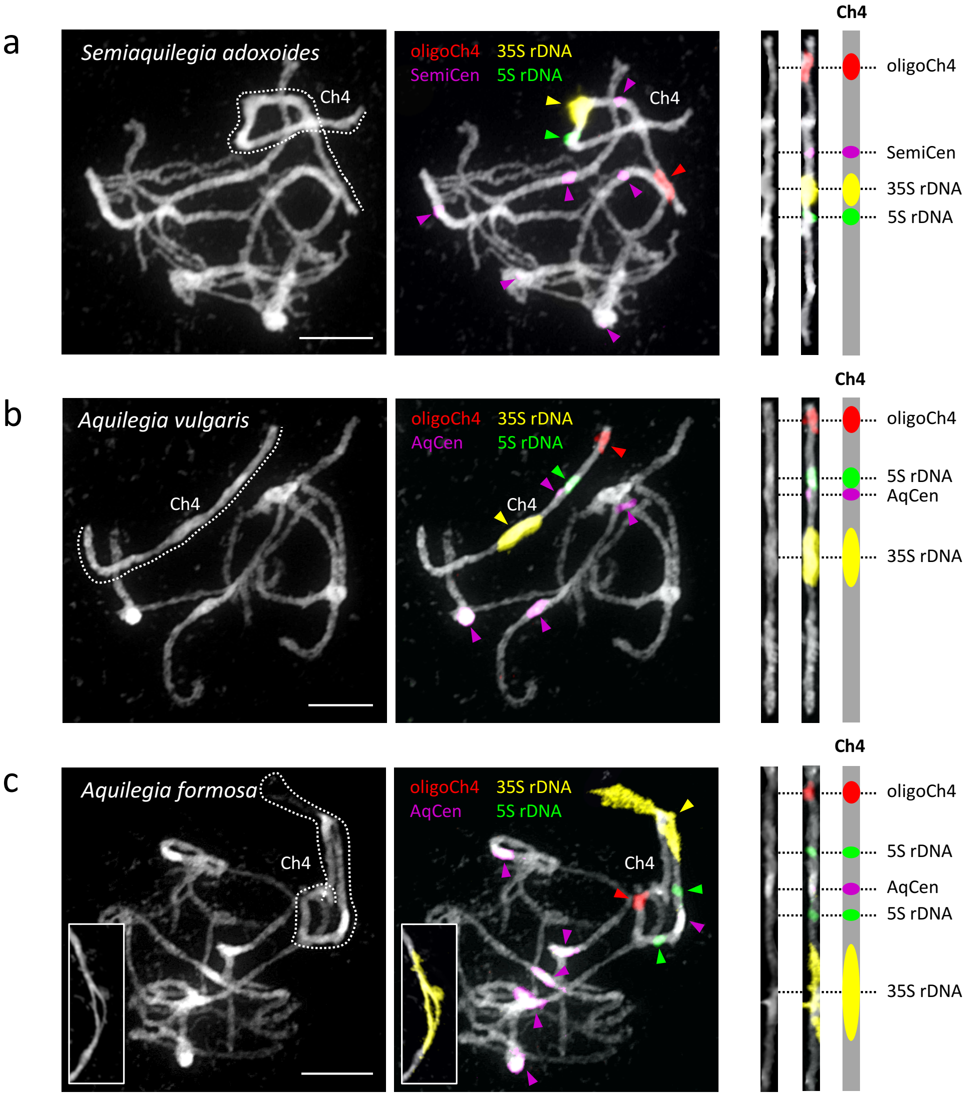


**Figure 7. Cytogenetic characterization of chromosome four in *Semiaquilegia* and *Aquilegia* species.** Pachytene chromosome spreads were probed with probes corresponding to oligoCh4 (red), 35S rDNA (yellow), 5S rDNA (green) and two (peri)centromeric tandem repeats (pink). Chromosomes were counterstained with DAPI. Scale bars = 10 µm.

## Discussion

We constructed a reference genome for the horticultural cultivar *Aquilegia coerulea* 'Goldsmith' and resequenced ten *Aquilegia* species with the goal of understanding the genomics of ecological speciation in this rapidly diversifying lineage. Variant sharing across the genus is widespread and deep, even across exceptionally large geographical distances. Although much of this sharing is presumably due to stochastic processes, as expected given the rapid time-scale of speciation, asymmetry of allele sharing demonstrates that the process of speciation has been reticulate throughout the genus, and that gene flow has been a common feature. *Aquilegia* species diversity therefore appears to be an example of ecological speciation, rather than being driven by the development of intrinsic barriers to gene flow[1,41,42]. In the future, studies incorporating more taxa and/or population-level variation will provide additional insight into the dynamics of this process. Given the extent of variant sharing, it will be also be interesting to explore the role of standing variation and admixture in adaptation throughout the genus.

Our analysis also led to the remarkable discovery that the evolutionary history of an entire chromosome differed from that of the rest of the genome. The average level of polymorphism on chromosome four is roughly twice that of the rest of the genome and gene trees on this chromosome appear to reflect a different species relationship (**Fig. 3**). Importantly, this chromosome is large and appears to be freely recombining, implying that these differences are unlikely to be due to a single evolutionary event, but rather reflect the accumulated effects of evolutionary forces acting differentially on the chromosome.

To the best of our knowledge, with the possible exception of sex chromosomes[43,44], such chromosome-wide patterns have never been observed before. Systematic differences between individual chromosomal regions have been described and attributed to the phenomenon known as "background selection", in which the interaction between recombination and purifying selection predicts a negative correlation between the local rate of recombination and coalescence. This interaction has, for example, lead to lower polymorphism in pericentromeric regions in many organisms. Although we observed no correlation between recombination and polymorphism in our data (**Supplementary Table 12**), the notably lower gene density on chromosome four suggests that background selection probably does play a role in elevating polymorphism on the chromosome.

Differences in gene content may thus be a proximal explanation for polymorphism levels on chromosome four, but we still lack a concrete explanation as to why these differences would have been established on chromosome four specifically. The roots of this go deep, extending at least into the genus *Semiaquilegia*. Although species with separate sexes exist in the Ranunculaceae, these transitions seem to be recent[45], and all *Aquilegia* and *Semiaquilegia* species are hermaphroditic. Furthermore, no heteromorphic sex chromosomes have been observed in the Ranunculales[46,47], making reverted sex chromosomes an unlikely hypothesis. It has also been suggested that chromosome four is a fusion of two homeologous chromosomes[39], as could result from the ancestral whole genome duplication[48–50], however, analysis of synteny blocks shows that this is not the case (Aköz, manuscript in preparation).

It is tempting to speculate that the distinct evolutionary history of chromosome four is connected to its large rDNA repeat clusters. Cytological[51,52] and phylogenetic[12,53,54] work separates the Ranunculaceae into two main subfamilies marked by different base chromosome numbers: the Thalictroideae (T-type, base n=7, including *Aquilegia* and *Semiaquilegia*) and the

Ranunculoideae (R-type, predominantly base n=8). In the three T-type species tested here, the 35S is proximal to the centromere, a localization seen for only 3.5% of 35S sites reported in higher plants[55]. In contrast, all R-type species examined have terminal or subterminal 45S loci[56–59]. Given that 35S repeats can be fragile sites[60] and 35S rDNA clusters and rearrangement breakpoints co-localize[61], a 35S-mediated chromosomal break could explain differences in base chromosome number between R-type and T-type species. Although we propose no mechanism whereby such rDNA-mediated changes could have lead to chromosome-wide decay, the rDNA cluster is, thus far, the only distinguishing physical feature of chromosome four. Comparative genomics work within the Ranunculaceae will therefore be useful for understanding the role that rDNA repeats have played in chromosome evolution and could provide additional insight into how rDNA could have contributed to chromosome four's variant evolutionary history.

In conclusion, the *Aquilegia* genus is a beautiful example of adaptive radiation through ecological speciation, and illustrates that standard population genetics models are not always sufficient to the explain the pattern of variation across the genome. A better understanding, and incorporation, of chromosome evolution is needed to give us a fuller picture of these evolutionary processes.

# Methods

## 1 Sequencing, Assembly, and Annotation

### 1.1 Sequencing

Sequencing was performed on *Aquilegia coerulea* cv 'Goldsmith', an inbred line constructed and provided by Todd Perkins of Goldsmith Seeds (now part of Syngenta). The line was of hybrid origin of multiple taxa and varieties of Aquilegia and then inbred. The sequencing reads were collected with standard Sanger sequencing protocols at the Department of Energy Joint Genome Institute in Walnut Creek, California and the HudsonAlpha Institute for Biotechnology. Libraries included two 2.5 Kb libraries (3.36x), two 6.5 Kb libraries (3.70x), two 33Kb insert size fosmid libraries (0.36x), and one 124kb insert size BAC library (0.17x). The final read set consists of 171,859 reads for a total of 3.411 Gb high quality bases (**Supplementary Table 15**).

### 1.2 Genome assembly and construction of pseudomolecule chromosomes

A total of 171,589 sequence reads (7.59x assembled sequence coverage) were assembled using our modified version of Arachne v.20071016[62] with parameters maxcliq1=120 n_haplotypes=2 max_bad_look=2000 START=SquashOverlaps BINGE_AND_PURGE_2HAP=True.

This produced 2,529 scaffolds (10,316 contigs), with a scaffold N50 of 3.1 Mb, 168 scaffolds larger than 100 kb, and total genome size of 298.6 Mb (**Supplementary Table 16**). Two genetic maps (*A. coerulea* 'Goldsmith' x *A. chrysantha* and *A. formosa* x *A. pubescens*) were used to identify 98 misjoins in the initial assembly. Misjoins were identified by a linkage group/syntenic discontinuity coincident with an area of low BAC/fosmid coverage. A total of 286 scaffolds were ordered and oriented with 279 joins to form 7 chromosomes. Each chromosome join is padded with 10,000 Ns. The remaining scaffolds were screened against bacterial proteins, organelle sequences, GenBank nr and removed if found to be a contaminant. Additional scaffolds were removed if they (a) consisted of >95% 24mers that occurred 4 other times in scaffolds larger than 50kb (957 scaffolds, 6.7 Mb), (b) contained only unanchored RNA sequences (14 scaffolds, 651.9 Kb), or (c) were less than 1kb in length (303 scaffolds). Significant telomeric sequence was identified using the TTTAGGG repeat, and care was taken to make sure that it was properly oriented in the production assembly. The final release assembly contains 1,034 scaffolds (7,930 contigs) that cover 291.7 Mb of the genome with a contig N50 of 110.9 kb and a scaffold L50 of 43.6 Mb (**Supplementary Table S17**.

## 1.3 Validation of genome assembly

Completeness of the euchromatic portion of the genome assembly was assessed using 81,617 full length cDNAs[22]. The aim of this analysis is to obtain a measure of completeness of the assembly, rather than a comprehensive examination of gene space. The cDNAs were aligned to the assembly using BLAT3 (Parameters: -t=dna –q=rna –extendThroughN -noHead) and alignments >=90% base pair identity and >=85% EST coverage were retained. The screened alignments indicate that 79,626 (98.69%) of the full length cDNAs aligned to the assembly. The cDNAs that failed to align were checked against the NCBI nucleotide repository (nr), and a large fraction were found to be arthropods (*Acyrthosiphon pisum*) and prokaryotes (*Acidovorax*).

A set of 23 BAC clones were sequenced in order to assess the accuracy of the assembly. Minor variants were detected in the comparison of the fosmid clones and the assembly. In all 23 BAC clones, the alignments were of high quality (< 0.35% bp error), with an overall bp error rate (including marked gap bases) in the BAC clones of 0.24% (1,831 discrepant bp out of 3,063,805; **Supplementary Table 18**).

## 1.4 Annotation

### *1.4.1 Genomic repeat and transposable element prediction*

Consensus repeat families were predicted de novo for the *A. coerulea* v3.1 genome by the RepeatModeler pipeline[63]. These consensus sequences were annotated for PFAM and Panther domains, and any sequences known to be associated with non-TE function were removed. The final curated library was used to generate a softmasked version of the *A. coerulea* v3.0 assembly.

### *1.4.2 Transcript assembly and gene model annotation*

A total of 246 million paired-end and a combined 1 billion single-end RNAseq reads from a diverse set of tissues and related *Aquilegia* species (**Supplementary Table 1**) were assembled using PERTRAN[64] to generate a candidate set containing 188,971 putative transcript assemblies. The PERTRAN candidate set was combined with 115,000 full length ESTs (the 85,000 sequence cDNA library derived from an *A. formosa X A. pubescens* cross[22] and 30,000 Sanger sequences of *A. formosa* sequenced at JGI) and aligned against the v3.0 release of the *A. coerulea* genome by PASA[65].

Loci were determined by BLAT alignments of above transcript assemblies and/or BLASTX of the proteomes of a diverse set of Angiosperms (*Arabidopsis thaliana* TAIR10, *Oryza sativa* v7, *Glycine max* Wm82.a2.v1, *Mimulus guttatus* v2, *Vitus vinifera* Genoscape.12X and *Poplar trichocarpa* v3.0). These protein homology seeds were further extended by the EXONERATE algorithm. Gene models were predicted by homology-based predictors, FGENESH+[66], FGENESH_EST (similar to FGENESH+, but using EST sequence to model splice sites and introns instead of putative translated sequence), and GenomeScan[67].

The final gene set was selected from all predictions at a given locus based on evidence for EST support or protein homology support according to several metrics, including Cscore, a protein BLASTP score ratio to homology seed mutual best hit (MBH) BLASTP score, and protein coverage, counted as the highest percentage of protein model aligned to the best of its Angiosperm homologs. A gene model was selected if its Cscore was at least 0.40 combined with protein homology coverage of at least 45%, or if the model had EST coverage of at least 50%. The predicted gene set was also filtered to remove gene models overlapping more than 20% with a masked Repeatmodeler consensus repeat region of the genome assembly, except for such cases that met more stringent score and coverage thresholds of 0.80 and 70% respectively. A final round of filtering to remove putative transposable elements was conducted using known TE PFAM and Panther domain homology present in more than 30% of the length of a given gene model. Finally, the selected gene models were improved by a second round of the PASA algorithm, which potentially included correction to selected intron splice sites, addition of UTR, and modeling of alternative spliceforms.

# 2 Sequencing of Species Individuals

## 2.1 Sequencing, mapping and variant calling

Individuals of 10 *Aquilegia* species and *Semiaquilegia adoxoides* were resequenced (**Fig. 1** and **Supplementary Table 19**. One sample (*A. pubescens*) was sequenced at the Vienna Biocenter Core Facilities Next Generation Sequencing (NGS) unit in Vienna, Austria and the others were sequenced at the JGI (Walnut Creek, CA, USA). All libraries were prepared using standard/slightly modified Illumina protocols and sequenced using paired-end Illumina sequencing. *Aquilegia* species read length was 100bp, the *S. adoxoides* read length was 150bp, and samples were sequenced to a depth of 58-124x coverage (**Supplementary Table 20**). Sequences were aligned against *A. coerulea* v3.1 with bwa mem (bwa mem -t 8 -p -M)[24,68]. Duplicates and unmapped reads were removed with SAMtools[23]. Picardtools[69] was used to clean the resulting bam files (CleanSam.jar), to remove duplicates (MarkDuplicates.jar), and to fix mate pair problems (FixMateInformation.jar). GATK 3.4[25,70] was used to identify problem intervals and do local realignments (RealignTargetCreator and IndelRealigner). The GATK Haplotype Caller was used to generate gVCF files for each sample. Individual gVCF files were merged and GenotypeGVCFs in GATK was used to call variants.

## 2.2 Variant filtration

Variants were filtered to identify positions in the single-copy genome that could be reliable called across all *Aquilegia* individuals. Variant Filtration in GATK 3.4[25,70] was used to filter multialleleic sites, indels +/-10bp, sites identified with Repeatmasker[71], and sites in previously-determined repetitive elements (see ***"Genomic repeat and transposable element prediction"*** *above*).

We required a minimum coverage of 15 in all samples and a genotype call (either variant or non-variant) in all accessions. Sites with less than 0.5x log median coverage or greater than -0.5x log median coverage in any sample were also removed. A table of the number of sites removed by each filter is in **Supplementary Table 21**.

### 2.3 Polarization

*S. adoxoides* was added to the *Aquilegia* individual species data set and the above filtration was repeated (**Supplementary Table 22**). The resulting variants were then polarized against *S. adoxoides*, resulting in nearly 1.5 million polarizable variant positions. A similar number of derived variants was detected in all species (**Supplementary Table 23**), suggesting no reference bias resulting from the largely North American provenance of the *A. coerulea* v.3.1 reference sequence used for mapping.

## 3 Evolutionary analysis

### 3.1 Basic population genetics

Basic population genetics parameters including nucleotide diversity (polymorphism and divergence) and $F_{ST}$ were calculated using custom scripts in R[72]. Nucleotide diversity was calculated as the percentage of pairwise differences in the mappable part of the genome. $F_{ST}$ was calculated as in Hudson *et al.*[73] To identify fourfold degenerate sites, four pseudo-vcfs replacing all positions with A,T,C, or G, respectively, were used as input into SNPeff[74] to assess the effect of each pseudo-variant in reference to the *A. coerulea* v3.1 annotation. Results from all four output files were compared to identify genic sites that caused no predicted protein changes.

### 3.2 Tree and cloudogram construction

Trees were constructed using a concatenated data set of all nonfiltered sites, either genome-wide or by chromosome. Neighbor joining (NJ) trees were made using the ape[75] and phangorn[76] packages in R[72] using a Hamming distance matrix and the nj command. RAxML trees were constructed using the default settings in RAxML[77]. All trees were bootstrapped 100 times. The cloudogram was made by constructing NJ trees using concatenated nonfiltered SNPs in non-overlapping 100kb windows across the genome (minimum of 100 variant positions per window, 2,387 trees total) and plotted using the densiTree package in phangorn[76].

### 3.3 Differences in subtree frequency by chromosome

For each of the 217 subtrees that had been observed in the cloudogram, we calculated the proportion of window trees on each chromosome containing the subtree of interest and performed a test of equal proportions (prop.test in R[72]) to determine whether the prevalence of the subtree varied by chromosome. For significantly-varying subtrees, we then performed another test of equal proportions (prop.test in R[72]) to ask whether subtree proportion on each chromosome was different from the genome-wide proportion. The appropriate Bonferroni multiple testing correction was applied to *p*-values obtained in both tests (n=217 and n=70, respectively).

### 3.4 Tests of D-statistics

D-statistics tests were performed in ANGSD[78] using non-filtered sites only. ANGSD ABBABABA was run with a block size of 100000 and results were bootstrapped. Tests were repeated using only fourfold degenerate sites.

### 3.5 Modelling effects of migration rate and effective population size

We simulated a simple coalescent model with the assumptions as follows: (1) population size is constant (N alleles) at all times, (2) *A. oxysepala* split from the population ancestral to *A. sibirica* and *A. japonica* at generation $t_2$=2*t, (3) *A. sibirica* and *A. japonica* split from each other at generation $t_1$=t, and (4) there was gene flow between *A. oxysepala* and *A. japonica* between t=0 and $t_1$. A first Markov Chain simulated migration with symmetric gene flow ($m_1$=$m_2$) and coalescence between t=0 and $t_1$ (*Five-State Markov Chain*, **Supplementary Table 24**). This process was run for T (t*N) generations to get the starting probabilities for the second process, which simulated coalescence between $t_1$ and $t_2$+1 (*Eight-State Markov Chain*, **Supplementary Table 25**). The second process was run for N generations. After first identifying a combination of parameters that minimized the difference between simulated versus observed gene genealogy proportions, we then reran the process with increased migration rate and/or N to check if simulated proportions matched observed chromosome 4-specific proportions. We also ran the initial chain under two additional models of gene flow: unidirectional ($m_2$=0) and asymmetric ($m_1$=2*$m_2$).

### 3.6 Robustness of chromosome four patterns to filtration

Variant filtration as outlined above was repeated with a stringent coverage filter (keeping only positions with +/-0.15x log median coverage in all samples) and nucleotide diversity per chromosome was recalculated. Nucleotide diversity per chromosome was also recalculated after removal of copy number variants detected by the readDepth package[79] in R[72], after the removal of tandem duplicates as determined by running DAGchainer[80] on *A. coerulea* v3.1 in CoGe SynMap[81], as well as after the removal of heterozygous variants for which both alleles were not supported by an equivalent number of reads (a log read number ratio <-0.3 or >0.3).

## 4 Construction of an *A. formosa* x *A. pubescens* genetic map

### 4.1 Mapping and variant detection

Construction of the *A. formosa* x *A. pubescens* F2 cross was previously described[82]. One *A. pubescens* F0 line (pub.2) and one *A. formosa* F0 line (form.2) had been sequenced as part of the species resequencing explained above. Libraries for the other *A. formosa* F0 (form.1) were constructed using a modified Illumina Nextera library preparation protocol[83] and sequenced at at the Vincent J. Coates Genomics Sequencing Laboratory (UC Berkeley). Libraries for the other *A. pubescens* F0 (pub.1), and for both F1 individuals (F1.1 and F1.2), were prepared using a slightly modified Illumina Genomic DNA Sample preparation protocol[84] and sequenced at the Vienna Biocenter Core Facilities Next generation sequencing (NGS) unit in Vienna, Austria. All individual libraries were sequenced as 100 bp paired-end reads on the Illumina HiSeq platform to 50-200x coverage. A subset of F2s were sequenced at the Vienna Biocenter Core Facilities Next Generation Sequencing (NGS) unit in Vienna, Austria (70 lines). Libraries for the remaining F2s (246 lines) were prepared and sequenced by the JGI (Walnut Creek, CA). All F2s were prepared using the Illumina multiplexing protocol and sequenced on the Illumina HiSeq platform to generate 100 bp paired end reads. Samples were 96-multiplexed to generate about 1-2x coverage. Sequences for all samples were aligned to the *A. coerulea* 'Goldsmith' v3.1 reference genome using bwa mem with default parameters[24,68]. SAMtools 0.1.19[85] mpileup (-q 60 -C 50 -B) was used to call variable sites in the F1 individuals. Variants were filtered for minimum base quality and minimum and maximum read depth (-Q 30, -d20, -D150) using SAMtools varFilter. Variable sites that had a genotype quality of 99 in the F1s were genotyped in F0 plants to generate a set of diagnostic alleles for each parent of origin. To assess nucleotide diversity, F0 and F1 samples were additionally processed with the mapping and variant calling pipeline as described for species samples above.

### 4.2 Genotyping of F2s and Genetic Map Construction

F2s were genotyped in genomic bins of 0.5 Mb in regions of moderate to high recombination and 1 Mb in regions with very low or no recombination, as estimated by the *A. coerulea* 'Goldsmith' x *A. chrysantha* cross used to assemble the *A. coerulea* 'Goldsmith' v3.1 reference genome (see "**Genome assembly and construction of pseudomolecule chromosomes**"). Ancestry of each bin was independently determined for each of the four parents. The ratio of:

$$\frac{\text{reads containing a diagnostic allele}}{\text{reads potentially containing a diagnostic allele}}$$

was calculated for each parent in each bin. If this ratio was $< 0.1$, the bin was assigned to the opposite parent and if the ratio was between 0.4 and 0.6, the bin was assigned to the parent containing the diagnostic allele. Otherwise, the bin was scored as missing data. Bin genotypes were used as markers to assemble a genetic map using R/qtl v.1.35-3[86]. To measure recombination in each F1 parent, genetic maps were initially constructed for each chromosome of a homolog pair in the F2s. After the two F1 homologous chromosome maps were estimated, data from each chromosome was combined to estimate the combined genetic map.

## 5 Chromosome four gene content and background selection

Unless noted, all analyses were done in R[72].

### 5.1 Gene content, Repeat content, and Variant Effects

Gene density and mean gene length were calculated considering primary transcripts only. Percent repetitive sequence was determined from annotation. The effects of variants was determined with SNPeff[74] using the filtered variant data set and primary transcripts.

### 5.2 Repeat family content

The RepeatClassifier utility from RepeatMasker[71] was used to assign *A . coerulea* v3.1 repeats to known repetitive element classes. For each of the 38 repeat families identified, the insertion rate per Mb was calculated for each chromosome and a permutation test was performed to determine whether this proportion was significantly different on chromosome four versus genome-wide. Briefly, we ran 1000 simulations to determine the number of insertions expected on chromosome four if insertions occurred randomly at the genome-wide insertion rate and then compared this distribution with the observed copy number in our data.

### 5.3 GO term enrichment

A two-sided Fisher's exact test was performed for each GO term to test whether the term made up a higher proportion on of genes on chromosome four versus the proportion in the rest of the genome. *P*-values were Bonferroni corrected for the number of GO terms (n=1936).

### 5.4 Background selection

A linear model was used to assess whether nucleotide diversity (polymorphism) in a 1MB window could be explained by gene density (bp genic/cM) within that window. Polymorphism was calculated as above, cM was determined from F2 analysis, and gene density (bp) was calculated using primary gene models from the *A. coerulea* v3.1 annotation.

## 6 Quantification of gene expression

We sequenced whole transcriptomes of sepals from 21 species of *Aquilegia* (**Supplementary Table S1**). Tissue was collected at the onset of anthesis and immediately immersed in RNAlater (Ambion) or snap frozen in liquid nitrogen. Total RNA was isolated using RNeasy kits (Qiagen) and mRNA was separated using poly-A pulldown (Illumina). Obtaining amounts of mRNA sufficient for preparation of sequencing libraries required pooling multiple sepals together into a single sample; we used tissue from a single individual when available, but often had to pool sepals from separate individuals into a single sample. We prepared sequencing libraries according to manufacturer's protocols except that some libraries were prepared using half-volume reactions (Illumina RNA-sequencing for *A. coerulea*, and half-volume Illumina TruSeq RNA for all other species). Libraries for *A. coerulea* were sequenced one sample per lane on an Illumina GAII (University of California, Davis Genome Center). Libraries for all other species were sequenced on an Illumina HiSeq at the Vincent J. Coates Genomics Sequencing Laboratory (UC Berkeley), with samples multiplexed using TruSeq indexed adapters (Illumina). Reads were aligned to *A. coerulea 'Goldsmith' v3.1* using bwa aln and bwa samse[68]. We processed alignments with SAMtools[85] and custom scripts were used to count the number of sequence reads per transcript for each sample. Reads that aligned ambiguously were probabilistically assigned to a single transcript. Read counts were normalized using calcNormFactors and cpm functions in the R package edgeR[87,88]. Mean abundance was calculated for each transcript by first averaging samples within a species, and then averaging across all species.

# 7 Cytology

## 7.1 Chromosome preparation

Inflorescences of the analyzed accessions were fixed in ethanol:acetic acid (3:1) overnight and stored in 70% ethanol at -20°C. Selected flower buds were rinsed in distilled water and citrate buffer (10 mM sodium citrate, pH 4.8; 2 × 5 min) and incubated in an enzyme mix (0.3% cellulase, cytohelicase, and pectolyase; all Sigma-Aldrich) in citrate buffer at 37°C for 3 to 6 h. Individual anthers were disintegrated on a microscope slide in a drop of citrate buffer and 15 to 30 µl of 60% acetic acid. The suspension was spread on a hot plate at 50°C for 0.5 to 2 min. Chromosomes were fixed by adding 100 µl of ethanol:acetic acid (3:1). The slide was dried with a hair dryer, postfixed in 4% formaldehyde dissolved in distilled water for 10 min, and air-dried. Chromosome preparations were treated with 100 µg/ml RNase in 2× sodium saline citrate (SSC; 20× SSC: 3 M sodium chloride, 300 mM trisodium citrate, pH 7.0) for 60 min and with 0.1 mg/ml pepsin in 0.01 M HCl at 37°C for 2 to 5 min; then postfixed in 4% formaldehyde in 2× SSC, and dehydrated in an ethanol series (70%, 90%, and 100%, 2 min each).

## 7.2 Probe preparation from oligo library

An oligonucleotide library consisting of 20,628 oligonucleotide probes to the 2 Mb region spanning the positions 42-44 Mbp of chromosome 4 (oligoCh4) was designed and synthesized by MYcroarray (Ann Arbor, MI). This library was used to prepare the chromosome 4-specific painting probe[40]. Briefly, oligonucleotides were multiplied in two independent amplification steps. First, 0.2 ng DNA from the immortal library was amplified from originally ligated adaptors in emulsion PCR using primers provided by MYcroarray together with the library. Emulsion PCR was used to increase the representativeness of amplified products[89]. Droplets were generated manually by stirring of oil phase at 1000 × g at 4°C for 10 min and then the aqueous phase was added. 500 ng of amplified product was used as a template for T7 in vitro transcription with MEGAshortscript T7 Kit (Invitrogen) – the second amplification step. RNA was purified on RNeasy spin columns (Qiagen) and labeled in reverse transcription with biotin-labeled R primer. The product – RNA:DNA hybrid – was washed using Zymo Quick-RNA MiniPrep (Zymo Research), hydrolysed by RNase and obtained DNA was cleaned again with Zymo Kit to get the final single-stranded DNA probe.

## 7.3 Fluorescence in situ hybridization (FISH)

Species resequencing data was used to determing the (peri)centromeric satellite repeats of *Semiaquilegia* (SemiCen) and *Aquilegia* (AqCen)[90]; the AqCen sequence corresponds to the previously-described centromeric repeat[90]. Bulked oligonucleotides specific for chromosome 4 (oligoCh4), (peri)centromeric satellite repeats, *Arabidopsis thaliana* BAC clone T15P10 (AF167571) containing 35S rRNA genes, and *A. thaliana* clone pCT4.2 (M65137) corresponding to a 500-bp 5S rRNA repeat were used as probes. All DNA probes were labeled with biotin-dUTP, digoxigenin-dUTP or Cy3-dUTP by nick translation[91] Selected labeled DNA probes were pooled together, ethanol precipitated, dissolved in 20 µl of 50% formamide, 10% dextran sulfate in 2× SSC and pipetted onto microscope slides. The slides were heated at 80°C for 2 min and incubated at 37°C overnight. Posthybridization washing was performed in 20% formamide in 2× SSC at 42°C (2 × 5 min). Hybridized probes were visualized through fluorescently labeled antibodies against biotin or digoxigenin[91]. Chromosomes were counterstained with 4',6-diamidino-2-phenylindole (DAPI, 2 µg/ml) in Vectashield antifade. Fluorescence signals were analyzed and photographed using a Zeiss Axioimager epifluorescence microscope and a CoolCube camera (MetaSystems). Individual images were merged and processed using Photoshop CS software (Adobe Systems). Pachytene chromosomes in Fig. 7 were straightened using the "straighten-curved-objects" plugin in the Image J software[92].

### 7.4 5-methylcytosine (5mC) immunodetection

For immunodetection of 5mC, chromosome spreads were prepared according to the procedure described above. Denaturation mixture containing 20 µl of 50% formamide, 10% dextran sulfate in 2× SSC was pipetted onto each microscope slide. The slides were heated at 80°C for 2 min, washed in 2× SSC (2 × 5 min) and incubated in bovine serum albumin solution (5% BSA, 0.2% Tween-20 in 4× SSC) at 37°C for 30 min. Immunodetection was performed using 100 µl of primary antibody against 5mC (mouse anti 5mC, Diagenode, diluted 1 : 100) at 37°C for 30 min. After washing in 2× SSC (2 × 5 min) the slides were incubated with the secondary antibody (Alexa Fluor 488 goat anti-mouse IgG, Invitrogen, diluted 1 : 200) at 37°C for 30 min, followed by washing in 2× SSC (2 × 5 min) and a dehydration in an ethanol series (70%, 90%, and 100%, 2 min each). Chromosomes were counterstained with DAPI, fluorescence signals analyzed and photographed as described above. The slides were washed in 2× SSC (2 × 5 min), dehydrated in an ethanol series (70%, 90%, and 100%, 2 min each), and rehybridized with 35S rDNA probe as described above.

## 8 Data availability

### 8.1 Species resequencing

*A. barnebyi* (SRRxxxxxx), *A. aurea* (SRR405095), *A. vulgaris* (SRR404349), *A. sibirica* (SRR405090), *A. formosa* (SRR408554), *A. japonica* (SRR413499), *A. oxysepala* (SRR413921), *A. longissima* (SRRXXXXXXX), *A. chrysantha* (SRR408559), *A. pubescens* (SRRXXXXxx (GMI - D14R)) are available in the Short Read Archive (https://www.ncbi.nlm.nih.gov/sra).

### 8.2 Whole genome *Aquilegia coerulea* 'Goldsmith'

Sanger sequences used for genome assembly are available in the NCBI Trace Archive (https://www.ncbi.nlm.nih.gov/Traces)

### 8.3 *Aquilegia coerulea* 'Goldsmith' ESTs

Available in the NCBI Short Read Archive (SRR505574-SRR505578)

### 8.4 *Aquilegia formosa* 412 ESTs

Available in the NCBI dbEST (https://www.ncbi.nlm.nih.gov/dbEST/)

### 8.5 *Aquilegia coerulea* 'Goldsmith' X *Aquilegia chrysantha* mapping population

Available in the NCBI Short Read Archive: SRRXXXXXX - SRRXXXXXXXX

### 8.6 *Aquilegia formosa* x *Aquilegia pubescens* mapping population

Available in the NCBI Short Read Archive : SRRXXXXXX - SRRXXXXXXX (JGI) and SRRXXXXXX - SRRXXXXXXX (GMI)

### 8.7 RNAseq

Available in the NCBI Short Read Archive: SRRXXXXXX - SRRXXXXXXX (UCSB)

### 8.8 Other files

vcfs of variants called in the *Aquilegia* species samples (both with and without *S. adoxoides*) are available for download at www.xxxxx.gmi.oeaw.ac.at.

## 9 URLs

The *A. coerulea* 'Goldsmith' v3.1 genome release is available at: https://phytozome.jgi.doe.gov/pz/portal.html

## References

**1.** Schluter, D. *The Ecology of Adaptive Radiation* (Oxford University Press, New York, 2000).

**2.** Hodges, S. A., Fulton, M., Yang, J. Y. & Whittall, J. B. Verne Grant and evolutionary studies of *Aquilegia*. *New Phytol.* **161**, 113–120 (2004).

**3.** Hodges, S. A. & Derieg, N. J. Adaptive radiations: from field to genomic studies. *Proc. Natl. Acad. Sci. United States Am.* **106 Suppl**, 9947–9954 (2009).

**4.** Kramer, E. M. *Aquilegia*: a new model for plant development, ecology, and evolution. *Annu. review plant biology* **60**, 261–277 (2009).

**5.** Munz, P. *Aquilegia*: The cultivated and wild columbines. *Gentes herbarum* **7**, 1–150 (1946).

**6.** Hodges, S. A. & Arnold, M. L. Floral and ecological isolation between *Aquilegia formosa* and *Aquilegia pubescens*. *Proc. Natl. Acad. Sci. United States Am.* **91**, 2493–2496 (1994).

**7.** Li, L. F. *et al.* Phenotypic and genetic evidence for ecological speciation of *Aquilegia japonica* and *A. oxysepala*. *New Phytol.* **204**, 1028–1040 (2014).

**8.** Taylor, R. J. Interspecific hybridization and its evolutionary significance in the genus *Aquilegia*. *Brittonia* **19**, 374 (1967).

**9.** Bastida, J. M., Alcántara, J. M., Rey, P. J., Vargas, P. & Herrera, C. M. Extended phylogeny of *Aquilegia*: The biogeographical and ecological patterns of two simultaneous but contrasting radiations. *Plant Syst. Evol.* **284**, 171–185 (2010).

**10.** Fior, S. *et al.* Spatiotemporal reconstruction of the *Aquilegia* rapid radiation through next-generation sequencing of rapidly evolving cpDNA regions. *New Phytol.* **198**, 579–592 (2013).

**11.** Whittall, J. B. & Hodges, S. A. Pollinator shifts drive increasingly long nectar spurs in columbine flowers. *Nat.* **447**, 706–709 (2007).

**12.** Ro, K. E. & Mcpheron, B. A. Molecular phylogeny of the *Aquilegia* group (Ranunculaceae) based on internal transcribed spacers and 5.8S nuclear ribosomal DNA. *Biochem. Syst. Ecol.* **25**, 445–461 (1997).

**13.** Novikova, P. Y. *et al.* Sequencing of the genus *Arabidopsis* identifies a complex history of nonbifurcating speciation and abundant trans-specific polymorphism. *Nat. Genet.* **48**, 1077–1082 (2016).

**14.** Svardal, H. *et al.* Ancient hybridization and strong adaptation to viruses across African vervet monkey populations. *Nat. Genet.* **49**, 1705–1713 (2017).

**15.** Malinsky, M. *et al.* Whole Genome Sequences Of Malawi Cichlids Reveal Multiple Radiations Interconnected By Gene Flow. *bioRxiv* 143859 (2017).

**16.** Takahata, N. Gene genealogy in three related populations: Consistency probability between gene and population trees. *Genet.* **122**, 957–66 (1989).

**17.** Avise, J. C. & Robinson, T. J. Hemiplasy: A new term in the lexicon of phylogenetics. *Syst. Biol.* **57**, 503–507 (2008).

**18.** Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Sci.* **328**, 710–722 (2010).

**19.** Lamichhaney, S. *et al.* Evolution of Darwin's finches and their beaks revealed by genome sequencing. *Nat.* **518**, 371–375 (2015).

**20.** Mallet, J., Besansky, N. & Hahn, M. W. How reticulated are species? *BioEssays* **38**, 140–149 (2016).

**21.** Pease, J. B., Haak, D. C., Hahn, M. W. & Moyle, L. C. Phylogenomics reveals three sources of adaptive variation during a rapid radiation. *PLoS Biol.* **14** (2016).

**22.** Kramer, E. & Hodges, S. *Aquilegia* as a model system for the evolution and ecology of petals. *Philos. transactions Royal Soc. Lond. Ser. B, Biol. sciences* **365**, 477–490 (2010).

**23.** Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma.* **25**, 2078–2079 (2009).

**24.** Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv:1303.3997v1 [q-bio.GN]* (2013).

**25.** McKenna, A. *et al.* The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

**26.** Cooper, E. A., Whittall, J. B., Hodges, S. A. & Nordborg, M. Genetic variation at nuclear loci fails to distinguish two morphologically distinct species of *Aquilegia*. *PLoS ONE* **5** (2010).

**27.** Leffler, E. M. *et al.* Revisiting an old riddle: what determines genetic diversity levels within species? *PLoS Biol.* **10** (2012).

**28.** Montalvo, A. M. Inbreeding depression and maternal effects in *Aquilegia caerulea*, a partially selfing plant. *Ecol.* **75**, 2395 (1994).

**29.** Herlihy, C. R. & Eckert, C. G. Genetic cost of reproductive assurance in a self-fertilizing plant. *Nat.* **416**, 320–323 (2002).

**30.** Yang, J. Y. & Hodges, S. a. Early Inbreeding depression selects for high outcrossing rates in *Aquilegia formosa* and *Aquilegia pubescens*. *Int. J. Of Plant Sci.* **171**, 860–871 (2010).

**31.** Loh, Y.-H. E. *et al.* Origins of shared genetic variation in African cichlids. *Mol. biology evolution* **30**, 906–17 (2013).

**32.** McVean, G. A. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nat.* **491**, 56–65 (2012).

**33.** Hodges, S. A. & Arnold, M. L. Columbines - a geographically widespread species flock. *Proc. Natl. Acad. Sci. United States Am.* **91**, 5129–5132 (1994).

**34.** Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).

**35.** Karasov, T. L., Horton, M. W. & Bergelson, J. Genomic variability as a driver of plant-pathogen coevolution? *Curr. Opin. Plant Biol.* **18**, 24–30 (2014).

**36.** Roux, F. & Bergelson, J. The genetics underlying natural variation in the biotic interactions of *Arabidopsis thaliana*: the challenges of linking evolutionary genetics and community ecology. *Curr. Top. Dev. Biol.* **119**, 111–156 (2016).

**37.** Nordborg, M. & Innan, H. The genealogy of sequences containing multiple sites subject to strong selection in a subdivided population. *Genet.* **163**, 1201–13 (2003).

**38.** Charlesworth, D., Charlesworth, B. & Morgan, M. T. The pattern of neutral molecular variation under the background selection model. *Genet.* **141**, 1619–32 (1995).

**39.** Linnert, G. Cytologische Untersuchungen an Arten und Artenbastarden von *Aquilegia* - I. Struktur und Polymorphismus der Nukleolen-chromosomen, Quadrivalente und B-chromosomen. *Chromosom.* **12**, 449–459 (1961).

**40.** Han, Y., Zhang, T., Thammapichai, P., Weng, Y. & Jiang, J. Chromosome-specific painting in Cucumis species using bulked oligonucleotides. *Genet.* **200**, 771–779 (2015).

**41.** Coyne, J. A., Orr, A. H. & Orr, H. A. *Speciation* (Sinauer, Sunderland, Massachusetts, 2004).

**42.** Seehausen, O. *et al.* Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192 (2014).

**43.** Toups, M. A. & Hahn, M. W. Retrogenes reveal the direction of sex-chromosome evolution in mosquitoes. *Genet.* **186**, 763–766 (2010).

**44.** Nam, K. *et al.* Extreme selective sweeps independently targeted the X chromosomes of the great apes. *Proc. Natl. Acad. Sci.* **112**, 6413–6418 (2015).

**45.** Soza, V. L., Brunet, J., Liston, A., Salles Smith, P. & Di Stilio, V. S. Phylogenetic insights into the correlates of dioecy in meadow-rues (*Thalictrum*, Ranunculaceae). *Mol. Phylogenetics Evol.* **63**, 180–192 (2012).

**46.** Westergaard, M. The mechanism of sex determination in dioecious flowering plants. *Adv. Genet.* **9**, 217–281 (1958).

**47.** Ming, R., Bendahmane, A. & Renner, S. S. Sex chromosomes in land plants. *Annu. Rev. Plant Biol.* **62**, 485–514 (2011).

**48.** Cui, L. *et al.* Widespread genome duplications throughout the history of flowering plants. *Genome Res.* **16**, 738–749 (2006).

**49.** Vanneste, K., Baele, G., Maere, S. & Van De Peer, Y. Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary. *Genome Res.* **24**, 1334–1347 (2014).

**50.** Tiley, G. P., Ané, C. & Burleigh, J. G. Evaluating and characterizing ancient whole-genome duplications in plants with gene count data. *Genome biology evolution* **8**, 1023–1037 (2016).

**51.** Langlet, O. Beiträge zur Zytologie der Ranunculazeen. *Svenk Bot. Tidskrift* **21**, 1–17 (1927).

**52.** Langlet, O. Über Chromosomenverhaltnisse und Systematik der Ranununculaceae. *Svenk Bot. Tidskrift* **26**, 1–2 (1932).

**53.** Wang, W., Lu, A. M., Ren, Y., Endress, M. E. & Chen, Z. D. Phylogeny and classification of Ranunculales: Evidence from four molecular loci and morphological data. *Perspectives Plant Ecol. Evol. Syst.* **11**, 81–110 (2009).

**54.** Cossard, G. *et al.* Subfamilial and tribal relationships of Ranunculaceae: evidence from eight molecular markers. *Plant Syst. Evol.* **302**, 419–431 (2016).

**55.** Roa, F. & Guerra, M. Distribution of 45S rDNA sites in chromosomes of plants: Structural and evolutionary implications. *BMC Evol. Biol.* **12**, 225 (2012).

**56.** Hizume, M., Shiraishi, H., Matsusaki, Y. & Shibata, F. Localization of 45S and 5S rDNA on chromosomes of *Nigella damascena*, Ranunculaceae. *CYTOLOGIA* **78**, 379–381 (2013).

**57.** Mlinarec, J., Papeš, D. A. & Besendorfer, V. Ribosomal, telomeric and heterochromatin sequences localization in the karyotype of *Anemone hortensis*. *Bot. J. Linnean Soc.* **150**, 177–186 (2006).

**58.** Weiss-Schneeweiss, H. *et al.* Chromosomal stasis in diploids contrasts with genome restructuring in auto- and allopolyploid taxa of *Hepatica* (Ranunculaceae). *New Phytol.* **174**, 669–682 (2007).

**59.** Liao, L. *et al.* Multiple hybridization origin of *Ranunculus cantoniensis* (4x): Evidence from trnL-F and ITS sequences and fluorescent in situ hybridization (FISH). *Plant Syst. Evol.* **276**, 31–37 (2008).

**60.** Huang, J., Ma, L., Yang, F., Fei, S. Z. & Li, L. 45S rDNA regions are chromosome fragile sites expressed as gaps in vitro on metaphase chromosomes of root-tip meristematic cells in Lolium spp. *PLoS ONE* **3** (2008).

**61.** Cazaux, B., Catalan, J., Veyrunes, F., Douzery, E. J. & Britton-Davidian, J. Are ribosomal DNA clusters rearrangement hotspots? A case study in the genus *Mus* (Rodentia, Muridae). *BMC Evol. Biol.* **11**, 124 (2011).

**62.** Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–96 (2003).

**63.** Smit, A. & Hubley, R. RepeatModeler v1.0.7 (2013). URL `www.repeatmasker.org/RepeatModeler/`.

**64.** Shu, S., Goodstein, D. & Rokhsar, D. PERTRAN: Genome-guided RNA-seq Read Assembler. - Report Number: LBNL-7081E (2013).

**65.** Haas, B. J. *et al.* Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).

**66.** Salamov, A. A. & Solovyev, V. V. Ab initio gene finding in Drosophila genomic DNA. *Genome Res.* **10**, 516–522 (2000).

**67.** Yeh, R. F., Lim, L. P. & Burge, C. B. Computational inference of homologous gene structures in the human genome. *Genome research* **11**, 803–16 (2001).

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

**69.** The Picard toolkit. URL http://broadinstitute.github.io/picard/.

**70.** DePristo, M. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* 491–498 (2011).

**71.** Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. 1996-2015.

**72.** R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2014). URL http://www.R-project.org/.

**73.** Hudson, R. R., Slatkin, M. & Maddison, W. P. Estimation of levels of gene flow from DNA sequence data. *Genet.* **132**, 583–9 (1992).

**74.** Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80–92 (2012).

**75.** Paradis, E., Claude, J. & Strimmer, K. APE: analyses of phylogenetics and evolution in R language. *Bioinforma.* **20**, 289–290 (2004).

**76.** Schliep, K. phangorn: phylogenetic analysis in R. *Bioinforma.* **27**, 592–593 (2011).

**77.** Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinforma.* **30**, 1312–1313 (2014).

**78.** Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinforma.* **15**, 356 (2014).

**79.** Miller, C. A., Hampton, O., Coarfa, C. & Milosavljevic, A. ReadDepth: A parallel R package for detecting copy number alterations from short sequencing reads. *PLoS ONE* **6**, e16327 (2011).

**80.** Haas, B. J., Delcher, A. L., Wortman, J. R. & Salzberg, S. L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinforma.* **20**, 3643–3646 (2004).

**81.** Lyons, E. & Freeling, M. How to usefully compare homologous plant genes and chromosomes as DNA sequences. *The Plant J.* **53**, 661–673 (2008).

**82.** Hodges, S. A., Whittall, J. B., Fulton, M. & Yang, J. Y. Genetics of floral traits influencing reproductive isolation between *Aquilegia formosa* and *Aquilegia pubescens*. *The Am. naturalist* **159 Suppl**, S51–S60 (2002).

**83.** Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10:e0128036** (2015).

**84.** Rabanal, F. A. *et al.* Unstable Inheritance of 45S rRNA Genes in Arabidopsis thaliana. *G3 (Bethesda, Md.)* **7**, 1201–1209 (2017).

**85.** Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinforma.* **25**, 2078–2079 (2009).

**86.** Broman, K. W., Wu, H., Sen, Ś. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinforma.* **19**, 889–890 (2003).

**87.** Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma.* 139–140 (2010).

**88.** McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

**89.** Murgha, Y. E., Rouillard, J.-M. & Gulari, E. Methods for the preparation of large quantities of complex single-stranded oligonucleotide libraries. *PLoS ONE* **9**, e94752 (2014).

**90.** Melters, D. P. *et al.* Comparative analysis of tandem repeats from hundreds of species reveals unique insights into centromere evolution. *Genome Biol.* **14**, R10 (2013).

**91.** Mandáková, T., Lysak, M. A., Mandáková, T. & Lysak, M. A. Painting of Arabidopsis Chromosomes with Chromosome-Specific BAC Clones. In *Current Protocols in Plant Biology*, 359–371 (John Wiley & Sons, Inc., Hoboken, NJ, USA, 2016).

**92.** Kocsis, E., Trus, B. L., Steer, C. J., Bisher, M. E. & Steven, A. C. Image averaging of flexible fibrous macromolecules: The clathrin triskelion has an elastic proximal segment. *J. Struct. Biol.* **107**, 6–14 (1991).

## Acknowledgements

The authors would like to thank Daniel Rokhsar for guidance in initiating the study and Todd Perkins of Goldsmith Seeds (now part of Syngenta) for providing *A. coerulea* 'Goldsmith' seeds. The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Next generation sequencing of some components of the *Aquilegia formosa* x *Aquilegia pubescens* mapping population was performed at the Vienna Biocenter Core Facilities Next Generation Sequencing (VBCF NGS) unit in Vienna, Austria (http://www.vbcf.ac.at). This work used the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, supported by NIH S10 Instrumentation Grants S10RR029668 and S10RR027303. T.M. and M.A.L. were supported by the Czech Science Foundation (grant no. P501/12/G090) and the CEITEC 2020 (grant no. LQ1601) project. M.K. was supported by National Program of Sustainability I (grant no. LO1204). E.S.B. was supported by the National Institutes of Health under the Ruth L. Kirschstein National Research Service Award (F32GM103154).

## Author contributions statement

M.N. and S.A.H. conceived the study; J.S., J.J., J.G., U.H., and K.B sequenced and assembled the reference; S.S. and. R.H. annotated the reference; G.A. performed coalescent-based modelling; E.B. performed F2 mapping; N.D. contributed RNAseq data and analysis; T.M and M.A.L. performed cytology; M.K. labeled the oligo paint library; J.Y., S.M, and V.N. constructed and sequenced libraries; D.F. analyzed species resequencing data, performed population genetics, and wrote the manuscript (with input from M.N., G.A., A.B., N.D. S.A.H., T.M., and M.A.L.); All authors read and approved the manuscript.

## Additional information

**Accession codes**

*A. barnebyi* - SRRxxxxxx

*A. aurea* - SRR405095

*A. vulgaris* - SRR404349

*A. sibirica* - SRR405090

*A. formosa* - SRR408554

*A. japonica* - SRR413499

*A. oxysepala* - SRR413921

*A. longissima* - SRRXXXXXX

*A. chrysantha* - SRR408559

*A. pubescens* - SRRXXXXXX

**Competing financial interests**.

The authors declare no competing interests.