The evolutionary dynamics of plant mating systems: how bias for studying 'interesting' plant reproductive systems could backfire

Elena M. Meyer[1], Laura F. Galloway[2], Andrew J. Eckert[1]

[1]Department of Biology, Virginia Commonwealth University, Richmond VA 23284, USA

[2]Department of Biology, University of Virginia, Charlottesville VA, 22904, USA

*Corresponding author: Elena M. Meyer, meyere3@vcu.edu

Abstract

An “abominable mystery”: angiosperm sexual systems have been a source of both interest and frustration for the botanical community since Darwin. The evolutionary stability, overall frequency, and distribution of self-fertilization and mixed-mating systems have been addressed in a variety of studies. However, there has been no recent study which directly addresses our knowledge of mating systems across families, the adequacy of existing data, or the potential for biases. Here, we present an updated dataset of mating systems across flowering plants covering 6,781 species and 212 families. We find that the vast majority of our data on mating systems comes from a small number of disproportionally-sampled families, and that families with significant proportions of dioecious or monoecious species are much more likely to be undersampled. This suggests that systematic study bias may mean we know less about this vital facet of plant life than we think.

Introduction

Darwin famously described the rapid emergence and evolutionary radiation of angiosperms as an “abominable mystery” (Darwin 1876, compiled 1903; Friedman 2009). One key to this mystery lies in the diverse reproductive strategies which angiosperms possess. Reproduction is a key determinant of fitness (Charlesworth 1994; Kalisz 1989), and thus it is a fundamental evolutionary driver. Despite this importance, our present understanding of the frequency and distribution of different reproductive strategies in angiosperms remains both limited and fraught with the potential for bias (Igic and Kohn 2006).

Plants possess an unusual level of variation in their reproductive strategies (Harder and Barrett 2006). Within angiosperms, reproduction can take place either asexually (via apomixis or by vegetative clonal growth) or sexually (2006). For sexual reproduction, two modes are possible: either self-fertilization (“selfing”) or outcrossing involving two parental plants. Many plant species, and even the same individuals within a species, perform both strategies, which is known as mixed mating (Stebbins 1950; Goodwille et al. 2005). In primarily outcrossing species, mechanisms often exist to avoid reproduction through selfing (e.g., genetic incompatibility with self pollen, anatomical separation of male and female flowers, or divergent phenological flowering time) (Lande and Schemske 1985). The diversity of mechanisms which plants use to prevent selfing suggests that outcrossing is often advantageous, and the avoidance of selfing has been associated with higher family diversification rates (Ferrer and Good 2012). However, the known phylogenetic breadth of selfing implies that despite these avoidance mechanisms, conditions exist where selfing can arise and persist, at least in the short-term.

Designation of sex in plants is also varied: individuals may be dioecious, with separate sexes corresponding to individuals; monoecious, with separate sexes corresponding to flowers on a single individual; or hermaphroditic, with both sexual organs present in the same floral structure. Self-fertilization is impossible for dioecious species due to their morphology. While selfing is possible for both monoecious and hermaphrodite species, in monoecy pollen must travel from the male to female flower for selfing to occur (i.e., geitonogamy). This requires either wind or pollinator facilitation. Dioecy and monoecy present an additional opportunity to enforce outcrossing through spatial or temporal separation of the sexes. Conversely, hermaphroditism could encourage selfing through proximity of the male and female parts of the flower. In this way, the ease of selfing for different sexual systems varies along a spectrum: selfing is impossible for dioecious species, more challenging for monoecious species, and most plausible for hermaphroditic species. Because of this correspondence, studies of plant sexual systems also provide information about mating systems. Mating systems and sexual systems are often addressed separately (for example, Goodwille et al. 2005 treats only mating systems, Wang et al. 2021 treats only sexual systems; but see Holsinger 2000). However, knowledge about sexual systems clearly informs knowledge about mating systems. Both are key traits that contribute to the unrivaled diversity of angiosperms (Barrett 2002), suggesting that additional synthesis of these traits will be beneficial to our understanding of reproduction in plants.

Considerable theoretical work has evaluated the evolutionary costs and benefits of selfing (Darwin 1876; Fisher 1941; Nagylacki 1976; Lloyd 1977, 1979, 1980; Cheptou 2019). Most straightforwardly, self-fertilization holds a transmission advantage of gene copies over two-parent reproduction (Fisher 1941). Based on this alone, without an evolutionary cost to selfing, it

would become commonplace. The obvious evolutionary cost to selfing is the increased expression of recessive deleterious phenotypes leading to inbreeding depression. Thus, to explain the variation we observe in plant reproductive systems, we must explain both the continued existence of selfing plants in the face of inbreeding depression and the persistence of outcrossing in spite of the two-fold transmission advantage of selfing. The fitness cost incurred by inbreeding depression (Charlesworth and Charlesworth 1987) may explain why selfing does not outcompete outcrossing based on transmission advantage alone (Lloyd 1979). Under circumstances where the cost of inbreeding depression does not outweigh the transmission advantage of selfing, however, selfing can still persist (Lloyd 1979). Another commonly cited advantage of selfing is reproductive assurance in the absence of a mate, or pollen limitation (Darwin 1876; Kalisz 2004; reviewed by Busch and Delph 2012). Selfing has thus also been proposed as a mechanism to aid colonization of new habitats, with self-compatible organisms more likely to be successful colonizers than self-incompatible species (Baker 1955). This is also sometimes referred to as "Baker's Law."

From this body of literature, two prominent – but not mutually exclusive – hypotheses emerged about the expected frequency of selfing across angiosperms. Stebbins (1957) proposed that selfing is an evolutionary "dead end," with selfing species more prone to extinction. He suggested that the limitations placed on genetic diversity by selfing would constrain adaptation to new environments. This hypothesis suggests, even if the deleterious alleles associated with inbreeding depression could be purged, that over time the loss of genetic diversity would doom selfing species to extinction. However, if transitions to selfing were sufficiently frequent, at any

given point in evolutionary time, we could still potentially observe extant selfing species (Stebbins 1957).

Schemske and Lande (1985) proposed that the distribution of selfing should be bimodal, with most species either primarily selfing or primarily outcrossing. Outcrossing would represent a stable strategy by avoiding inbreeding depression and promoting diversity through biparental sexual reproduction. Similarly, highly selfing species would be able to avoid inbreeding depression by purging deleterious alleles and would thus persist due to their transmission advantage and reproductive assurance. Species with mixed mating, however, should be less common, as purging of deleterious alleles would be more difficult under intermediate levels of selfing, yet the risks of inbreeding depression and loss of genetic diversity remain.

Since Schemske and Lande (1985), the distribution of selfing across seed plants has been revisited multiple times. Vogler and Kalisz (2001) found that the bimodal distribution hypothesis was supported for wind-pollinated species, but not for animal-pollinated species. This was revised again when Goodwille et al. (2005) reviewed mixed mating (defined in their review by an outcrossing rate, *t*, falling between 0.2 and 0.8). They showed that mixed-mating systems are more frequent than previously thought (in their study, up to 42% of species selfed or had the potential to self), and the frequency at which mixed-mating is observed to occur suggested some level of stability for selfing. The authors additionally cautioned, however, that their data was “far from random,” and drew 44% of its total species from only five families (Goodwillie et al. 2005). Shortly after, Igic and Kohn (2006) demonstrated a bias against obligate outcrossing species in published studies of outcrossing and selfing rates and cautioned that any generalizations about the true distribution of selfing were likely premature.

All of these studies (Schemske and Lande 1985, Vogler and Kalisz 2001, Goodwille et al. 2005) explicitly address the question of the true underlying distribution of selfing. Since the publication of these seminal papers, the rich and provocative theoretical background surrounding selfing has promoted a wealth of empirical papers addressing selfing in various plant systems (for many good examples, see Whitehead et al. 2018). This has greatly expanded the dataset of plant species with known mating systems. Schemske and Lande (1985) presented a dataset of just 55 species to demonstrate the bimodal distribution empirically. Now, known mating systems number in the thousands, as shown in this review. However, this larger dataset has the potential to continue promoting the existing biases noted by Igic and Kohn (2006).

Existing studies fall into one of three categories: (1) papers which directly address selfing rate evolution and selfing frequency across multiple plant families, (2) papers which address the interplay of the selfing trait with other topics in ecology and evolution, or (3) studies that focus on specific families which are known to have varied or potentially labile mating systems. A focus on a few families with especially varied mating systems is often necessary for individual studies. But over time, this will inevitably lead to nonrandom and biased sampling across the angiosperm phylogeny with respect to mating systems, since the focus is on selfing species and not producing random samples across the angiosperm phylogeny (*cf.* Igic and Kohn 2006).

Papers which address the interplay of selfing at a multi-family level with other various topics in ecology and evolution have been especially prolific. For example, selfing has been explored in conjunction with geographic range dynamics (Grossenbacher et al. 2015, 2016), invasive status and naturalization (Rananjatovo et al. 2016), island biogeography (Grossenbacher 2017), within-population selfing rate variation (Whitehead et al. 2018), life history and range

size (Prior and Busch 2021), and seed size (Tateyama et al. 2021). Additionally, studies focused on estimating rates of transitions to and from selfing using phylogenetic comparative methods (PCMs) have contributed significantly to knowledge about the tempo and mode of selfing evolution. Such studies have provided detailed knowledge about the role of selfing to diversification rates in Solanaceae (Goldberg et al. 2015), Onagraceae (Freyman and Hoha 2019), and Fabaceae (Delaney and Igic 2021), but these families by definition were selected because they have labile mating systems. While this is appropriate for the goals set forth in each study, this will likely reinforce existing biases against families with high frequencies of outcrossing. Thus, our collective knowledge about the underlying distribution of selfing, if based on meta-analyses using these studies, will likely be biased.

Numerous reviews focusing more broadly on selfing also exist (e.g., Goodwille et al. 2005, Wright et al. 2013). However, the question of the underlying distribution of self-compatibility (and, conversely, self-sterility) has not been recently revisited. In other words, we remain extremely uncertain about what percentage of angiosperm species are capable of selfing. This means we also do not know to what extent our data allow us to make appropriate estimation of trends across angiosperm-wide diversity. In previous work, the significant proportion of known mating systems which incorporate selfing has been used to suggest the stability of mixed mating (Goodwille et al. 2005). However, if selfing and mixed-mating are selectively over-sampled, this assumption could be premature. Similarly, if our understanding of the true underlying distribution of selfing is biased, then we run the risk of making incorrect and overly simplifying assumptions when modeling mating system evolution with use of PCMs. Incorrect or

biased assumptions, moreover, may more generally lead us to incorrect conclusions, even when estimating quantities as simple as the frequency of selfing and mixed mating across angiosperms.

The evolutionary dynamics of plant mating systems are more than just a challenging theoretical problem. They also quantify fundamental driving forces determining fitness in real plant populations. Allocation of selfing and outcrossing across individuals varies widely within populations and is labile in populations responding to environmental change (Whitehead et al. 2018). Theory suggests that selfing may be adaptive in rapidly changing environments by providing reproductive assurance, especially in situations of pollinator limitations (Darwin 1876, Baker 1955, Cheptou 2019). However, selfing may also limit adaptive potential, especially in the long term (Stebbins 1957). In an era of rapid anthropogenic climate change, some species have already changed their reproductive allocation in response to climate (Rech 2021; Austin 2022, Cheptou et al. 2022). Similarly, significant pollinator decline has been a hallmark of our era (Thomman et al. 2013), which can lead to pollen limitation (Ashman et al. 2004). This could lead to evolutionary pressures promoting the rapid change of reproductive traits to adapt (Thomman et al. 2013). Thus, in the Anthropocene, selfing could be favored in the short-term by climate instability but lead to extinction in the longer term as the known costs to extensive selfing are realized. However, some species may have stress-induced mechanisms to promote outcrossing as opposed to selfing during episodes of drastic environmental change (van Ginkel and Flipphi 2020). It is thus paramount to better understand the existing biases in our knowledge about the underlying distribution of plant mating systems in our rapidly changing environments.

Unraveling these complex dynamics requires an adequate foundation of fundamental biological knowledge about plant reproduction, such as the proportion of species which can self-

fertilize and how these species are distributed across the angiosperm phylogeny. To this end, we present a dataset reviewing the mating systems of 6,781 unique species, including a total of 212 families. Of these species, 1,270 are newly synthesized from single-species studies or small reviews of under 100 species. Using these data, we explicitly address and examine potential underlying biases in the data and identify families which are over or under-sampled compared to their overall contributions to angiosperm diversity. We examined the role of existing knowledge about plant sexual systems (dioecy, monoecy, and hermaphroditism) in shaping potential sampling biases. Broadly, we revisit the question of the underlying distribution of selfing, the adequacy of our existing data, review historical change, and propose future best practices. We then extend these results to discuss bias when quantifying the frequency of selfing across angiosperms. Our results suggest that systematic study bias favoring 'interesting' mating systems like selfing and mixed-mating impacts our understanding of mating systems in angiosperms, meaning that our understanding of this vital facet of plant life is more limited than previously thought.

Literature Review and Scope

We conducted an extensive literature review using Web of Science, utilizing the following topic search string: "TS=(('breeding system' OR 'mating system' OR 'self-compatib*' OR 'self-fertil*' OR autogam* OR 'auto-fertil*' OR outcross* OR apomixis) AND plant)". We chose these varied search terms to better capture all possible studies containing mating system information, as terminology varies significantly in plant reproductive biology (Cardoso et al. 2018). Additional filters were used to choose results from relevant journals and exclude journals

which were not relevant (e.g., medical journals). The full search strings that were utilized for the review are provided in file **S1.** Journals were included from January 2017 until September 2022. This timeframe was chosen based on the existence of multiple reviews which we judged as covering existing work from the pre-2017 timeframe adequately (e.g., Goodwille et al. 2005, Moeller 2017; see **Table 1,** which summarizes previously-conducted reviews).

The filtered search produced 6,334 results. These were manually reviewed based on their abstracts for applicability. From these results, we found 1,618 studies with potentially relevant data. These studies were reviewed in full, and included if they provided, at a minimum, a binary designation of either self-compatible (SC) or self-incompatible (SI) for the focal species. In keeping with previous studies (Goodwille et al. 2005, Moeller et al. 2017), papers which reported a sexual system (i.e., apomixis) but not a selfing designation were excluded. Sexual relatives of apomictic species are generally selfing, and apomixis has been hypothesized to be correlated with selfing (Hörandl 2009, Prior and Busch 2021), so this was unlikely to impact the inclusion of outcrossing species in the dataset. Additionally, crop species were excluded, as breeding efforts and crop domestication often directly select on the ability to self (Solís Montero et al. 2021). Uncultivated crop wild relatives, which are not subjected to similar pressures, were included. Names for all species were standardized to the World Flora Online (https://www.worldfloraonline.org/), using the R package "U.Taxonstand" (R version 4.3.0, R Core Team 2023) and the WFO dataset provided by the authors (Zhang and Qian 2022). Following this review process, we included 761 studies from the 2017-2022 period. This process is fully summarized in **Figure 1,** illustrating how we processed and synthesized both existing reviews and new data into a unified dataset.

This literature review also included 11 large review studies (*n* > 100 species) where selfing data were also identified (Table 1). These studies were not only limited to the 2017-2022 timeframe, as they were included to represent and cover pre-2017 knowledge. These are largely synthesized in their entirety, but as some of these reviews included non-angiosperm seed plants, we excluded all non-angiosperm species (e.g., gymnosperms) for the sake of parity. Notably, while Ferrer and Good (2011) was reviewed and presented a large dataset focused on self-incompatibility/self-sterility, their work was not included in this study because the data were not accessible (i.e., numerous attempts were made to contact both authors at various time points, as well as the editorial staff at the journal in which this paper was published, but we were ultimately unsuccessful in obtaining the raw data).

Based on our synthesis of the review data (Table 1), and the 761 additional studies from 2017-2022 we newly identified from the literature, we were able to identify a total of 6,781 unique species with known mating systems. This includes 1,270 newly synthesized species. These 6,781 species draw from 212 families, or approximately half of all known families recognized by APG IV (The Angiosperm Phylogeny Group, 2016). To our knowledge, this represents the largest known single dataset on plant reproductive systems to date.

To better understand bias in sampling of mating systems, we also leveraged the higher availability of information on sexual systems. Approximately 23% of angiosperm species have a known sexual system (Wang et al. 2021). We used this as a point of reference to estimate frequencies of dioecy, monoecy, and hermaphroditism by family, based on calculating the frequency of observed species with each sexual system over the observed number of species for the family. The number of species in each family was based on Christenhusz and Byng (2016), which follows APG IV (The Angiosperm Phylogeny Group, 2016). Based on the families we

identified with elevated dioecy and monoecy, we assessed representation in the mating system literature under the hypothesis that non-hermaphroditic systems would be less likely to be included in the mating system dataset. We also assessed if over-represented families had enriched hermaphroditism compared to other angiosperm families. This was done using a randomization test, in which we iteratively sampled across all families with known sexual systems (n = 9999 samples of 10 randomly chosen families, including over-represented families) for average percent hermaphroditism. We then compared the distribution of the randomly-sampled mean hermaphroditism with that of over-represented family groups.

To assess and visualize evenness of sampling between families, we utilized a method related to rank-abundance curves from ecology. Each family in the dataset was ranked based on the total number of representative species in the dataset, and then the number of species represented was normalized to the family with the highest number of species. We also examined how this changed over time using the reviews of Schemske and Lande (1985), Goodwille et al. (2005), and Moeller (2017). All datasets were filtered to include only angiosperm species prior to analysis.

Results and Discussion

Few families have well-studied mating systems proportional to overall diversity within families

Our goal was to better understand the state of knowledge of plant reproductive systems across all angiosperms. To cover the pre-2017 period, we synthesized major reviews of selfing, starting with Schemske and Lande (1985). To cover the 2017-2022 period, we reviewed 761 more recent papers from the 2017-2022 period, comprising 6,781 unique species. In total, this

species list has at least one representative from 217 families, or approximately half (52.2%) of all known angiosperm families (Christenhusz and Byng 2016). To the best of our knowledge, this is the largest such dataset presented to date, both with respect to the number of species, and the total number of families represented.

From our dataset (excluding species with conflicting designations), 62.8% of species were SC and thus selfed or had the potential to self at least some of the time. This includes both obligate selfing species and mixed-mating species. An additional 38.5% were SI and hence obligately outcrossing. Species with multiple conflicting designations (i.e., studies indicating both SC and SI for the same species) were excluded. This decision was based on similar practices in other mating system studies (Moeller et al. 2017, Prior and Busch 2021). These species with conflicting designations comprised 7.5% of the dataset. In contrast to our results, Goodwille et al. (2005) reported 42% of their sample as mixed-mating. While our proportion of SC species was larger, our sample includes SC species which would have been excluded from the mixed-mating sample of Goodwillie et al. (2005) based on their definition of mixed mating. Igic and Kohn's (2006) discussion of potential biases in mating system studies proposed a corrected 48.7% frequency of SI based on their estimation of under-sampling of obligate outcrossing species.

While our review synthesized a large number of studies, known trait data for selfing remains limited compared to overall angiosperm diversity. Additionally, sampling is highly uneven across family groups. Thus, despite the size of our dataset, many families have a limited coverage of taxonomically described species. For example, out of the 212 families identified here, only 51 families had more than 5% of their total species represented, while only 34 families had more than 10% of their total species represented. Additionally, only 14 of these 34 families

had more than 100 species covered, suggesting limited understanding of larger family groups. To illustrate these trends, the coverage for larger (n > 500 sp.) families is summarized in **Table 2**. This shows how many large families have limited representation, and a small number of families are significantly better studied than others.

Indeed, a disproportionate amount of the overall dataset comes from a limited number of well-covered families which differs drastically from these families' contribution to overall angiosperm diversity. If species were randomly sampled across the angiosperm phylogeny, these percent coverages should more closely match. For example, species classified within Solanaceae represent 7.7% of our data, but with its approximately 2,600 species covered at 20.04% in our dataset, this family represents only 0.9% of total estimated angiosperm species. Similarly, Asteraceae, Fabaceae, Brassicaceae, Onagraceae, and Bromeliaceae are all overrepresented in our dataset compared to their expected proportions based on family size if random sampling was used as the criteria to select species for inclusion in our dataset. Some of the largest families, moreover, are notably underrepresented: Orchidaceae, Rubiaceae, and Poaceae (**Figure 2).**

It is important to note that our overall dataset represents a very small portion of the overall diversity of flowering plants. There are 295,383 known plant species recognized in APG IV (Christenhusz and Byng 2016), meaning that out of all species, only approximately 2% have a documented mating system in our dataset. For comparison, plant sexual systems (i.e., monoecious, dioecious, or hermaphrodite) are known for at least 23% of all species in APG IV (Wang et al. 2021). While determining the mating system of a species can be more challenging than determining the sexual system, this is nonetheless an example of a related floral trait for which greatly more information is available.

Existing data arises disproportionally from a small number of families with labile mating systems

A small number of families comprised the majority of existing plant mating system data. The top 10 best-covered families in our review contained over half of the total species (50.90%). In contrast, these 10 families contribute only 1.2% of overall angiosperm diversity. Similarly, if we look at the top 10 families in our review by raw number of species with known mating systems (instead of percent coverage), these top 10 families represent 64.7% of the total species in the review, but still only 33.51% of total angiosperm biodiversity. We calculated rank-abundance plots for Schemske and Lande (1985), Goodwille et al. (2005), Moeller et al. (2017), and our study to quantify trends in sampling dominance over time (**Figure 3**). For comparison, these are shown alongside the expected angiosperm-wide rank-abundance curve using current estimates of angiosperm diversity and taxonomy. As time progresses, the rank-abundance curves for sampled species more closely match the expectation given our current taxonomic knowledge. However, as shown in **Figure 4**, the most-abundant families which make up the steep top of the abundance curves are not the same families as expected. If sampling was truly random across angiosperms, Orchidaceae, Asteraceae, Fabaceae, Rubiaceae, and Poaceae – the most abundant families in our current understanding of angiosperm diversity – would form the top of the curve in all instances, especially those later in time. This is clearly not the case. Furthermore, the contributions of these families to our collective understanding of plant mating systems varies over time and across studies to such an extent that it is questionable whether summaries of existing data are unbiased.

Some of the unevenness in sampling across families likely arises from studies focused on using PCMs to describe evolutionary transitions to and from selfing. It is notable that several of

the best-studied families (Table 2) have also been the focus of PCM studies (Solanaceae, Goldberg et al. 2010; Onagraceae, Freyman and Hoha 2019; Fabaceae, Igic and Delaney 2022). One can also imagine this interaction in the opposite way; these families were chosen for PCM study due to the greater data availability. Regardless of the direction of this interaction, the inevitable result is that our understanding of the evolution, ecological characteristics, and frequency of selfing arises disproportionately from a limited number of families that may not be representative of angiosperms overall.

Choosing families with labile mating systems is appropriate when using PCMs to estimate rates of evolutionary transitions to and from selfing, as well as the correlations of these transitions with overall diversification rates. Due to better availability of trait data at tree tips, statistical power for parameter estimation using PCMs is maximized for these families (Harmon 2018). In Onagraceae, for example, autogamy has long been known to evolve easily, and has evolved at least 150 times independently (Raven 1979). Similarly, Solanaceae is known to have intermediate levels of self-incompatibility and frequent mating system transitions (Goldberg et al. 2010, 2012). The only other family with similarly high coverage with respect to mating system data is Phrymaceae, which contains the genus *Mimulus* (previously classified in the now defunct Scrophulariaceae), a frequent model system for studies of selfing and ecological genomic studies (Wu et al. 2008). Perhaps unsurprisingly, in our dataset, all representatives of Phrymaceae come from the genus *Mimulus.* Studies of selfing in *Mimulus* could explain why more mating system information exists for this particular family, as it has become a model system for many different questions in ecology and evolution. In a similar way, the field of genetics is grappling with model organism choice in an era of increased availability of genomic data, with some arguing that existing model organisms are insufficient to capture the diversity

and complexity of the systems they are supposed to represent (Bertile et al. 2023). The study of plant reproductive systems likely faces a similar problem: families which are used as model systems, and thus contribute disproportionately to available knowledge on mating systems, are not representative of angiosperm diversity as a whole.

Families which have an elevated frequency of dioecious or monoecious species are less likely to be well-studied

Considering the uneven sampling shown above, we also utilized more-available information about plant sexual systems (categorized here as monoecy, dioecy, and hermaphroditism, although some more specialized mating systems such as polygamy do not fit neatly into these categories) to examine potential biases in mating system studies. This is sensible given the clear biological links between sexual systems and mating systems. For instance, if a bias existed against families with a significant number of dioecious species, this would suggest a bias against obligate-outcrossing species. Similarly, self-fertilization in monoecious species could be more challenging than in their hermaphroditic counterparts, and under-sampling of families with known high levels of monoecy could also suggest bias based on preexisting knowledge about sexual system and morphology.

For dioecy, the relationship between sexual system and mating system is straightforward: dioecious plants are morphologically and definitionally unable to self. Dioecy is comparatively uncommon in sexual systems (Renner and Ricklefs, 1995), but some families have considerably enriched dioecy (n = 15 families with > 200 species). These families had between 10%-49% dioecious species (Wang et al. 2021). From these highly dioecious families, all 15 rank in the bottom 50% of families for mating system coverage in our dataset, and all but two rank in the bottom 30%. Additionally, one third of these families (Smilacaceae, Cucurbitaceae, Vitaceae,

Putranjivaceae, and Menispermaceae) did not have a single species with a known mating system listed in any published study (**Figure 5**). This is especially surprising as Cucurbitaceae and Vitaceae are families with economically important species.

The relationship between the mating system and sexual system for monecious species is more complex. While monoecy is frequently hypothesized to be an adaptation to limit selfing, empirical evidence is limited: Bertin (1993) theorized if monoecy was an adaptation to prevent selfing, SI would be uncommon in species exhibiting this sexual system. Unexpectedly, Bertin (1993) found monoecy to be independent of the mating system. However, for our data, highly monoecious families (defined as families with >5% monoecy) were also underrepresented in our dataset relative to their expectation based on overall angiosperm diversity, and we found that the 15 families with the greatest levels of known monoecy fall well below the mean level of coverage in the plant mating systems data (Table 4). As a corollary to this, families with high levels of dioecious and monoecious species typically have very limited hermaphroditic species. While only an incomplete sample of total species have a known sexual system, most families that are exhibiting high dioecy or monoecy also have very limited or no hermaphroditism. The percent hermaphroditism for each family is shown in **Tables 3 and 4**, which shows that underrepresentation of monoecy and dioecy likely also yielded an enrichment in hermaphroditism.

We examined the families which were over-represented with respect to selfing information (**Figure 2**) and found these families to be enriched for hermaphroditism (**Figure 6**). The mean hermaphroditism for families in Wang et al. (2021) was 72%, and the mean for Solanaceae, Asteraceae, Fabaceae, Brassicaceae, Onagraceae, and Bromeliaceae was 87%. To better quantity this deviation from the expected mean, we conducted an ad-hoc randomization

test. We iterated 9999 random draws of six families from the 330 families with known sexual systems in Wang et al. (2021) and generated a frequency distribution of mean hermaphroditism. Our mean hermaphroditism (72%) for over-represented families can be seen to fall to the right tail of the distribution (**Figure 6**). This suggests that not only are dioecious and monoecious species under-sampled, but hermaphroditic species are over-sampled. This is consistent with a general bias against outcrossing species, which leads to a positive bias for selfing species. Thus, existing data, even when expanded to sexual systems, likely overestimates the frequency of selfing species.

Data about plant sexual systems are also incomplete. However, persistent undersampling of dioecious and monoecious families points to a larger pattern of under-representation of outcrossing in mating and sexual system studies. These results suggest that species which are known to have limited capacity for selfing due to their anatomy are less likely to be included in databases about plant mating systems. Furthermore, as selfing is correlated with a wide variety of other traits – the so-called "selfing syndrome" (Darwin, 1879, Ornduff 1969) – this could directly impact our understanding of other traits in angiosperms using the same resources as presented here.

Based on our survey of mating system literature, we initially produced an estimate of 62.5% SC and 38.5% SI. However, in light of the biases we discuss above, we must consider that these estimates over-estimate SC, and under-estimate SI. Based on our synthesis of sexual system data, we could leverage this information to make some informed choices when accounting for bias in mating system data. Based on Wang et al. (2021), dioecy accounts for about 8% of total plant species, monoecy 16%, and hermaphroditism 76%. While the family-level under and oversampling suggests a larger issue of focus on SC and mixed mating, a simple

initial correction could be to correct for limited dioecy. For species we studied with both mating system and sexual system data (n=2217), our sample has about 2% dioecy, 13% monoecy, and 85% hermaphroditism. If we were to correct for under-representation of SI contributions from dioecy and monoecy (assuming dioecious species are always SI, and 63% of monoecy species are SI, based on our data), we would have a new estimate of 58% SC and 42% SI. But as noted previously, the patterns of sampling bias which produced the observed deficiency of dioecy and monoecy in mating system data could also impact sampling in hermaphroditic species. Thus it seems entirely plausible that our existing estimates remain SC-enriched.

Conclusions and Future Directions

By assembling existing knowledge of selfing in an updated, harmonized, and publicly available dataset, we hope to promote further research in the rich field of plant reproductive systems. We especially hope to encourage studies which address the underlying distribution of selfing in angiosperms, and selfing-focused papers broadly to consider the adequacy of existing data. We also hope to promote additional investigation into as-of-yet understudied groups, such as understudied families with dioecious and monoecious species, or families which have been shown to be especially under-represented compared to their species diversity.

As we have shown, obligate outcrossing species may be systematically underrepresented, which was previously shown by Igic and Kohn (2006). We have additionally shown that this bias is likely informed by prior knowledge about families, leading to “interesting” mating systems like selfing and mixed mating being over-studied. Additionally, a small number of families represent a large amount of existing knowledge about plant mating systems. Some of these families may contain a higher-than-average number of selfing species based on enriched

hermaphroditism. In order to address this problem, we need not only *more* data, but thoughtfully collected data. For example, data which is collected randomly or which focuses on improving knowledge of understudied groups would be especially useful at this juncture. As illustrated by our work, it will also be vital that we utilize a more holistic view of plant reproduction. Sexual systems and other mating-system associated traits which are more available can be leveraged to help us generate reasonable estimates of selfing or to assess the data that we *do* have. For example, authors studying mating systems within specific family groups would be well-advised to examine family-level data about frequency of SI, SC, and different sexual systems within their family, and to consider critically the

Our results suggest that we understand less about the underlying frequency of selfing across angiosperms than was previously thought. Self-fertilization presents a potential route for reproductive assurance for plants in challenging and unstable environments (Baker 1955, Busch and Delph 2012), but can have significant determinants as well. In an era of unprecedented anthropogenic disturbance, it thus seems plausible that an individual plant's reproductive strategy could become a stronger predictor of its fitness than it has been in the past. Reproductive systems are known to be labile in many large families and across populations within species (Whitehead et al. 2018).

It could be imagined that under anthropogenic climate change, selfing could be the way of the future as reproductive assurance is needed for short-term survival of some species. However, in the long term, the known deleterious consequences of inbreeding could limit fitness or lead to extinction in selfing species. This could represent an as-of-yet unforeseen cost of anthropogenic change on angiosperm diversity. This could even illuminate a new feedback loop of biodiversity loss in plants, which are a foundational piece of all terrestrial ecosystems. This

illustrates the importance of understanding the prevalence of outcrossing, selfing and mixed-mating across angiosperms.

Citations

Ashman, T. L., Knight, T. M., Steets, J. A., Amarasekare, P., Burd, M., Campbell, D. R., … Wilson, W. G. (2004). Pollen limitation of plant reproduction: Ecological and evolutionary causes and consequences. *Ecology*. Ecological Society of America. 85(9), 2408-2421. https://doi.org/10.1890/03-8024

Baker, H. G. (1955). Self-compatibility and establishment after "long-distance" dispersal. *Evolution*, *9*(3), 347. https://doi.org/10.2307/2405656

Barrett, S. C. H. (2002). The evolution of plant sexual diversity. *Nature Reviews Genetics*, 3(4), 274-284. https://doi.org/10.1038/nrg776

Bertile, F., Matallana-Surget, S., Tholey, A., Cristobal S., Armengaud J. (2023). Diversifying the concept of model organisms in the age of -omics. *Commun Biol* 6(1), 1062. https://doi.org/10.1038/s42003-023-05458-x

Busch, J. W., & Delph, L. F. (2012). The relative importance of reproductive assurance and automatic selection as hypotheses for the evolution of self-fertilization. *Annals of Botany*, 109(3), 553–562. https://doi.org/10.1093/aob/mcr219

Cardoso, J. C. F., Viana, M. L., Matias, R., Furtado, M. T., Caetano, A. P. de S., Consolaro, H., & Brito, V. L. G. de. (2018). Towards a unified terminology for angiosperm reproductive system. *Acta Botanica Brasilica*. Sociedade Botanica do Brasil. 32(3), 329-348. https://doi.org/10.1590/0102-33062018abb0124

Charesworth B., 1994. Evolution in age-structured populations. Cambridge University Press, Cambridge: 10.1017/CBO9780511525711

Charlesworth, D., & Charlesworth, B. (1987). Inbreeding depression and its evolutionary consequences. Annual Review of Ecology and Systematics, 18(1), 237-268. https://doi.org/10.1146/annurev.es.18.110187.001321

Chase, M. W., Christenhusz, M. J. M., Fay, M. F., Byng, J. W., Judd, W. S., Soltis, D. E., … Weber, A. (2016). An update of the Angiosperm Phylogeny Group classification for the orders

and families of flowering plants: APG IV. Botanical Journal of the Linnean Society, 181(1), 1–20. https://doi.org/10.1111/boj.12385

Cheptou, P. O. (2019). Does the evolution of self-fertilization rescue populations or increase the risk of extinction? *Annals of Botany*. Oxford University Press. 123(2), 337-345. https://doi.org/10.1093/aob/mcy144

Cheptou, P. O., Imbert, E., & Thomann, M. (2022). Rapid evolution of selfing syndrome traits in Viola arvensis revealed by resurrection ecology. *American Journal of Botany*, *109*(11), 1838–1846. https://doi.org/10.1002/ajb2.16028

Darwin C. Letter 395. To J.D. Hooker. Down, July 22nd, 1879. In: Darwin F, Seward AC, editors. *More Letters of Charles Darwin.* London, Murray; 1903.

Darwin C. (1876). The effects of cross and self-fertilization in the vegetable kingdom. London, Murray.

Fisher, R.A. (1941). Average excess and average effect of a gene substitution. *Annals of Eugenics*, 11(1), 53-63. https://doi.org/10.1111/j.1469-1809.1941.tb02272.x

Friedman, W. E. (2009). The meaning of Darwin's "abominable mystery." *American Journal of Botany*. 96(1), 5-21. https://doi.org/10.3732/ajb.0800150

Goldberg, E. E., & Igić, B. (2012). Tempo and mode in plant breeding system evolution. *Evolution*, *66*(12), 3701–3709. https://doi.org/10.1111/j.1558-5646.2012.01730.x

Goldberg, E. E., Kohn, J. R., Lande, R., Robertson, K. A., Smith, S. A., & Igić, B. (2010). Species selection maintains self-incompatibility. *Science*, 330(6003), 493-495.https://doi.org/10.1126/science.1194513

Goodwillie, C., Kalisz, S., & Eckert, C. G. (2005). The evolutionary enigma of mixed mating systems in plants: Occurrence, theoretical explanations, and empirical evidence. In *Annual Review of Ecology, Evolution, and Systematics.* 36, 47-79. https://doi.org/10.1146/annurev.ecolsys.36.091704.175539

Grossenbacher, D., Briscoe Runquist, R., Goldberg, E. E., & Brandvain, Y. (2015). Geographic range size is predicted by plant mating system. *Ecology Letters*, 18(7), 706–713. https://doi.org/10.1111/ele.12449

Grossenbacher, D., Briscoe Runquist, R. D., Goldberg, E. E., & Brandvain, Y. (2016). No association between plant mating system and geographic range overlap. *American Journal of Botany*, 103(1), 110–117. https://doi.org/10.3732/ajb.1500078

Grossenbacher, D. L., Brandvain, Y., Auld, J. R., Burd, M., Cheptou, P. O., Conner, J. K., … Goldberg, E. E. (2017). Self-compatibility is over-represented on islands. *New Phytologist*, 215(1), 469–478. https://doi.org/10.1111/nph.14534

Harder, L.D., Barrett, S.C.H., (2016). The Ecology and Evolution of Flowers. Oxford Univ Press, Oxford.

Harmon, J. H. (2018). Phylogenetic Comparative Methods. Version 1.1. https://lukejharmon.github.io/pcm/

Holsinger, K. E. (2000). Reproductive systems and evolution in vascular plants. *Proceedings of the National Academy of Sciences of the United States of America*. 97(13), 7037-7042. https://doi.org/10.1073/pnas.97.13.7037

Hörandl E. (2010). The evolution of self-fertility in apomictic plants. *Sexual Plant Reproduction*, 23(1), 73–86. https://doi.org/10.1007/s00497-009-0122-3

Igic, B., & Kohn, J. R. (2006). The distribution of plant mating systems: study bias against obligately outcrossing species. *Evolution*, 60(5), 1098. https://doi.org/10.1554/05-383.1

Kalisz, S. (1989). Fitness consequences of mating system, seed weight, and emergence date in a winter annual, Collinsia verna. *Evolution*, 43(6). https://doi.org/10.1111/j.1558-5646.1989.tb02573.x

Kalisz, S., Vogler, D. W., & Hanley, K. M. (2004). Context-dependent autonomous self-fertilization yields reproductive assurance and mixed mating. Nature, 430(7002), 884–887. https://doi.org/10.1038/nature02776

Lloyd, D. G. (1977). Genetic and phenotypic models of natural selection. *Journal of Theoretical Biology*, *69*(3), 543–560. https://doi.org/10.1016/0022-5193(77)90155-2

Lloyd, D. G. (1979). Some reproductive factors affecting the selection of self-fertilization in plants. *The American Naturalist*, 113(1), 67–79. https://doi.org/10.1086/283365

Lloyd, D. G. (1980). Demographic factors and mating patterns in angiosperms. In O. T. Solbrig (Ed.), Lloyd, D. G. (1980). Demographic factors and mating patterns in angiosperms. I(Vol. 15, pp. 67–88). Blackwell.

Nagylaki T. (1976). A model for the evolution of self-fertilization and vegetative reproduction. *Journal of Theoretical Biology*, 58(1), 55–58. https://doi.org/10.1016/0022-5193(76)90138-7

Owens, S. J., & Miller, R. (2009). Cross- and self-fertilization of plants - Darwin's experiments and what we know now. *Botanical Journal of the Linnean Society* 161(4), S75-S113. https://doi.org/10.1111/j.1095-8339.2009.01010.x

Prior, C. J., & Busch, J. W. (2021). Selfing rate variation within species is unrelated to life-history traits or geographic range position. *American Journal of Botany*, 108(11), 2294–2308. https://doi.org/10.1002/ajb2.1766

Razanajatovo, M., van Kleunen, M., Kreft, H., Dawson, W., Essl, F., Pergl, J., … Weigelt, P. (2019). Autofertility and self-compatibility moderately benefit island colonization of plants. *Global Ecology and Biogeography*, 28(3), 341–352. https://doi.org/10.1111/geb.12854

Rech AR, Ollerton J, Dalsgaard B, et al. (2021). Population-level plant pollination mode is influenced by Quaternary climate and pollinators. *Biotropica*. 53(2), 632–642. https://doi.org/10.1111/btp.12905

Renner, S. S., & Ricklefs, R. E. (1995). Dioecy and its correlates in the flowering plants. American Journal of Botany, 82(5), 596. https://doi.org/10.2307/2445418.

Solís-Montero, L., Aceves-Chong, L., Vega-Polanco, M., & Vargas-Ponce, O. (2021). Changes in reproductive traits in Physalis philadelphica; an unexpected shift toward self-incompatibility in a domesticated annual fruit crop. Frontiers in Plant Science, 12. https://doi.org/10.3389/fpls.2021.658406

Stebbins, G. L. (1950). Variation and evolution in plants. New York, Columbia University Press.

Tateyama, H., Chimura, K., & Tsuchimatsu, T. (2021). Evolution of seed mass associated with mating systems in multiple plant families. *Journal of Evolutionary Biology*, *34*(12), 1981–1987. https://doi.org/10.1111/jeb.13949

Thomann, M., Imbert, E., Devaux, C., & Cheptou, P. O. (2013. Flowering plants under global pollinator decline. *Trends in Plant Science*. 18(7), 353-359. https://doi.org/10.1016/j.tplants.2013.04.002

Whitehead, M. R., Lanfear, R., Mitchell, R. J., & Karron, J. D. (2018). Plant mating systems often vary widely among populations. *Frontiers in Ecology and Evolution*. 6(38), https://doi.org/10.3389/fevo.2018.00038

Wright, S. I., Kalisz, S., & Slotte, T. (2013). Evolutionary consequences of self-fertilization in plants. *Proceedings of the Royal Society B: Biological Sciences*. 280(1760). https://doi.org/10.1098/rspb.2013.0133

Wu, C. A., Lowry, D. B., Cooley, A. M., Wright, K. M., Lee, Y. W., & Willis, J. H. (2008). Mimulus is an emerging model system for the integration of ecological and genomic studies. *Heredity*. 100(2), 220-230. https://doi.org/10.1038/sj.hdy.6801018

Zhang, J., & Qian, H. (2023). U.Taxonstand: An R package for standardizing scientific names of plants and animals. *Plant Diversity*, *45*(1), 1–5. https://doi.org/10.1016/j.pld.2022.09.001

Tables:

Table 1. Large reviews (n>100) providing data on plant mating systems (n=11 reviews).

Year	Dataset	Entries (n)	Includes
1985	Schemske and Lande	55	
2005	Goodwille et al.	345	1985
2015, 2016	Grossenbacher et al.	588	
2016	Razanajatovo et al.	1890	
2017	Moeller et al.	624	1985, 2005
2018	Whitehead et al.	741	
2019	Freyman and Hohna	357	
2021	Delaney and Igic	1410	
2021	Tateyama et al	642	2015/2016 (partial)
2021	Prior and Busch	772	

Table 2. Within-family coverage obtained in our review of plant mating systems, for families of 500 species or more with 3% or greater coverage in our dataset.

Family	Species (n)	Coverage (%)
Onagraceae	656	48.63
Solanaceae	2600	20.04
Brassicaceae	3628	9.40
Lythraceae	620	9.19
Geraniaceae	830	8.43
Fabaceae	19500	6.96
Bignoniaceae	870	4.37
Bromeliaceae	3475	4.35
Balsaminaceae	1000	3.90
Proteaceae	1660	3.80

Asteraceae	24700	3.62
Plantaginaceae	1900	3.53
Dipterocarpaceae	695	3.45
Cactaceae	1750	3.37

Table 3. Families with enriched dioecy based on Wang et al. (2021), by percent of known dioecious species relative to overall family size. Percent of family known to be hermaphroditic is also shown.

Family	**Species (n)**	**Dioecy (%)**	**Coverage (%)**	**Herm (%)**
Smilacaceae	255	49.02	0.00	0.00
Aquifoliaceae	500	44.80	0.60	0.00
Salicaceae	1220	40.00	1.31	0.00
Pentaphylacaceae	330	27.58	0.00	26.4
Menispermaceae	440	25.00	0.00	0.00
Dioscoreaceae	715	21.12	0.28	3.82
Burseraceae	615	19.02	0.52	13.1
Cucurbitaceae	965	17.51	0.52	0.00
Santalaceae	1000	16.90	0.90	53.9
Ebenaceae	800	16.13	0.13	0.00
Moraceae	1180	14.07	0.59	0.00
Actinidiaceae	360	12.78	0.00	40.4
Vitaceae	910	11.76	0.22	22.1
Putranjivaceae	216	11.11	0.00	0.00
Loranthaceae	1050	9.71	1.43	85.9

Table 4. Families with enriched monoecy based on Wang et al. (2021), by percent of known monoecious species relative to overall family size. Percent of family known to be hermaphroditic is also shown.

Family	**Species (n)**	**Monoecy (%)**	**Coverage (%)**	**Herm (%)**
Fagaceae	927	69.04	3.02	0.00
Begoniaceae	1825	56.55	0.55	0.00
Euphorbiaceae	6252	23.22	0.99	<0.1
Cyperaceae	5500	22.84	0.36	37.86
Moraceae	1180	19.41	0.59	0.00
Cucurbitaceae	965	16.06	0.52	0.00

Asteraceae	24700	13.59	3.62	42.0
Santalaceae	1000	11.30	0.90	53.9
Amaranthaceae	2040	8.82	0.44	48.9
Phyllanthaceae	2050	8.29	0.15	2.47
Araceae	3750	7.28	0.43	38.9
Poaceae	12000	6.08	1.77	76.3
Urticaceae	2625	6.02	0.11	<0.1
Sapindaceae	1860	5.16	0.65	7.3
Eriocaulaceae	1207	5.05	<0.01	0.00

Include as file for S1:

Additional search for journals which were relevant to the literature review and timeframe: "and **Agronomy** or **Plant Sciences** or **Ecology** or **Genetics Heredity** or **Evolutionary Biology** or **Horticulture** or **Multidisciplinary Sciences** or **Environmental Sciences** or **Biology** or **Biodiversity Conservation**or **Forestry** or **Reproductive Biology** or **Developmental Biology** or **Environmental Studies** or **History Philosophy Of Science** or **Anatomy Morphology** (Web of Science Categories) and **2017** or **2018** or **2019** or **2020** or **2021** or **2022** (Publication Years)".

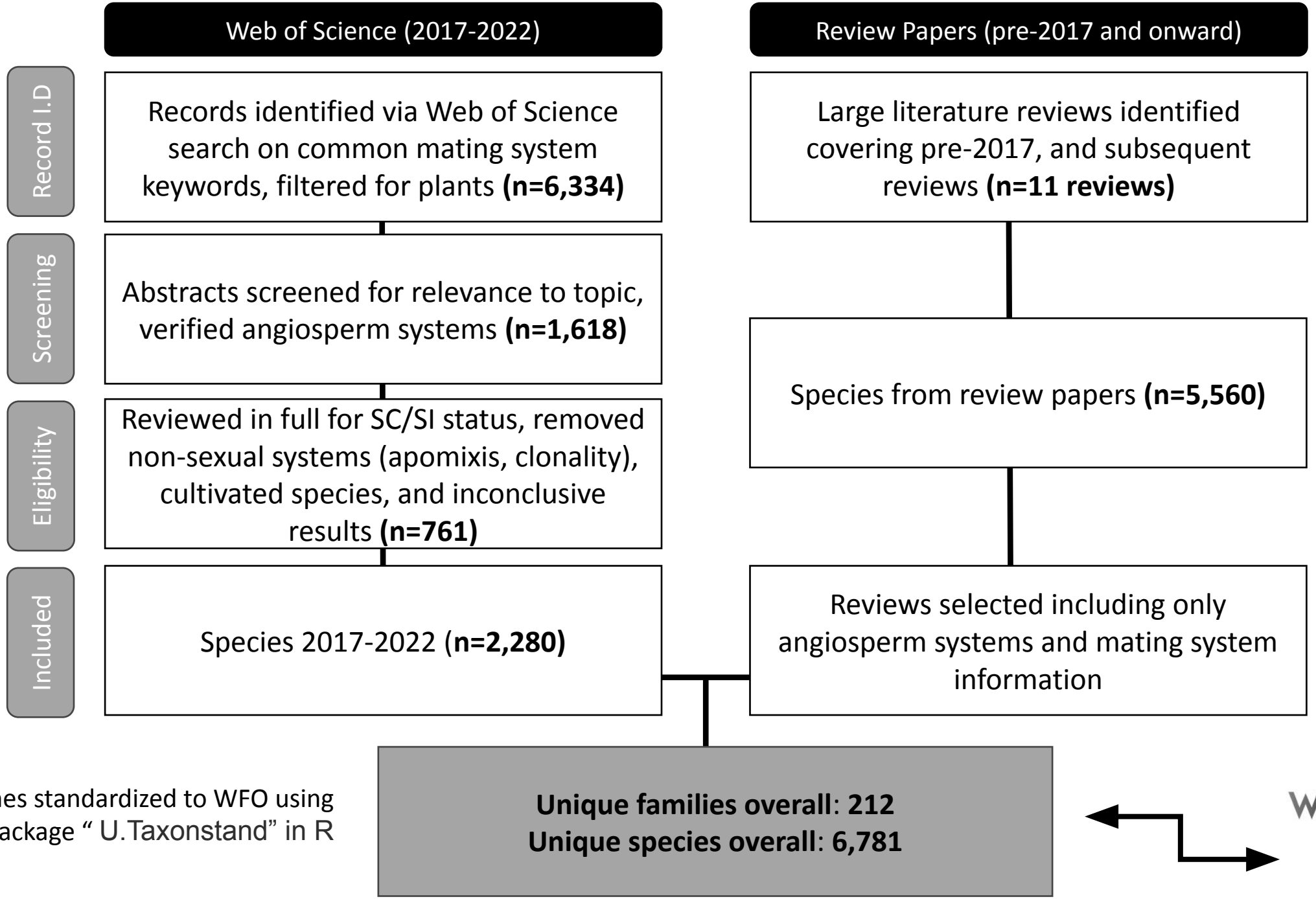

Figure 1. Literature review flow chart.

**Names standardized to WFO using package " U.Taxonstand" in R

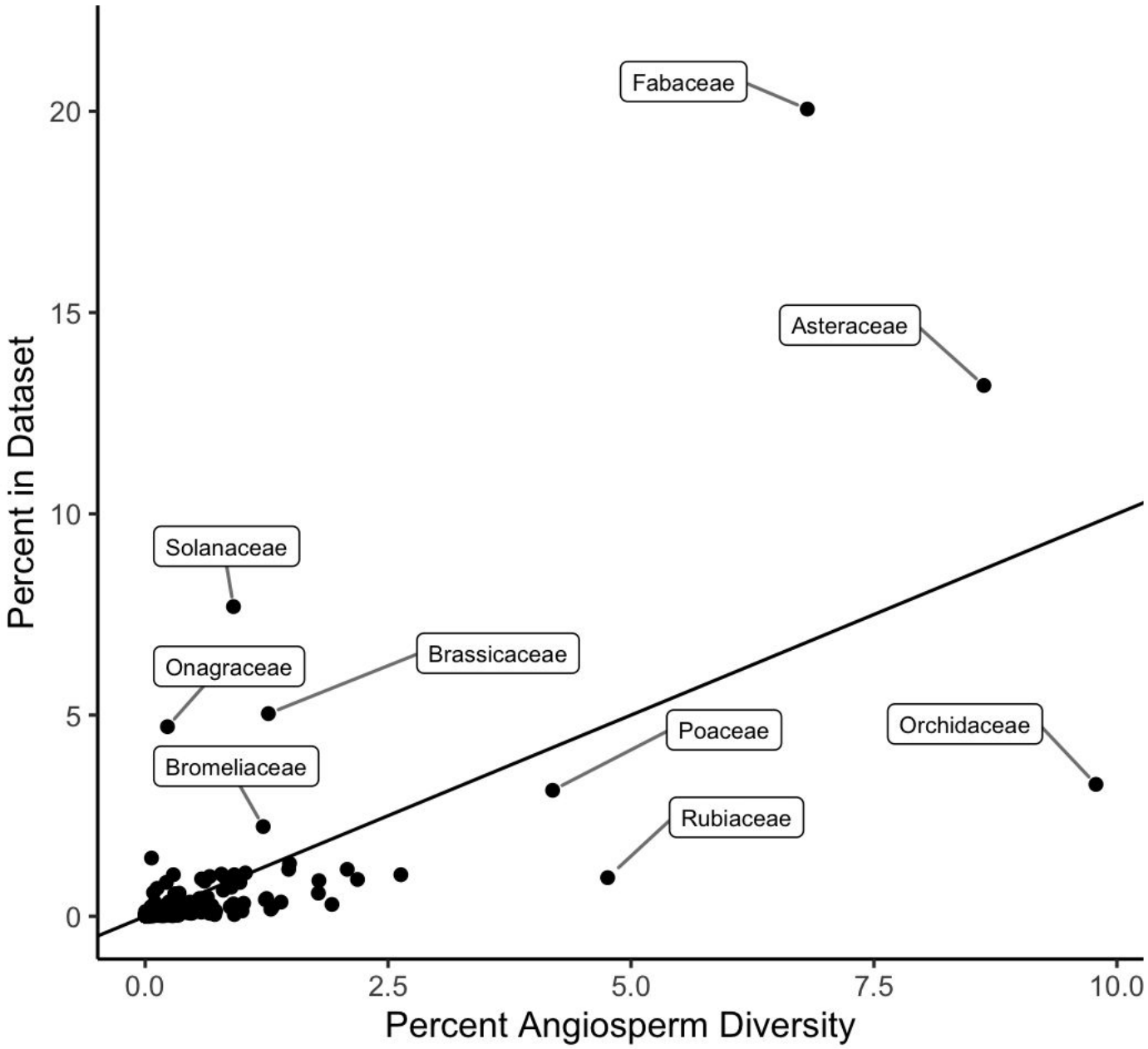

Figure 2. Families represented in our mating system dataset plotted by the percent of angiosperm diversity they represent (x-axis), and the percent of our dataset they represent. Points falling above the 1:1 line are over-represented, points falling below it are under-represented.

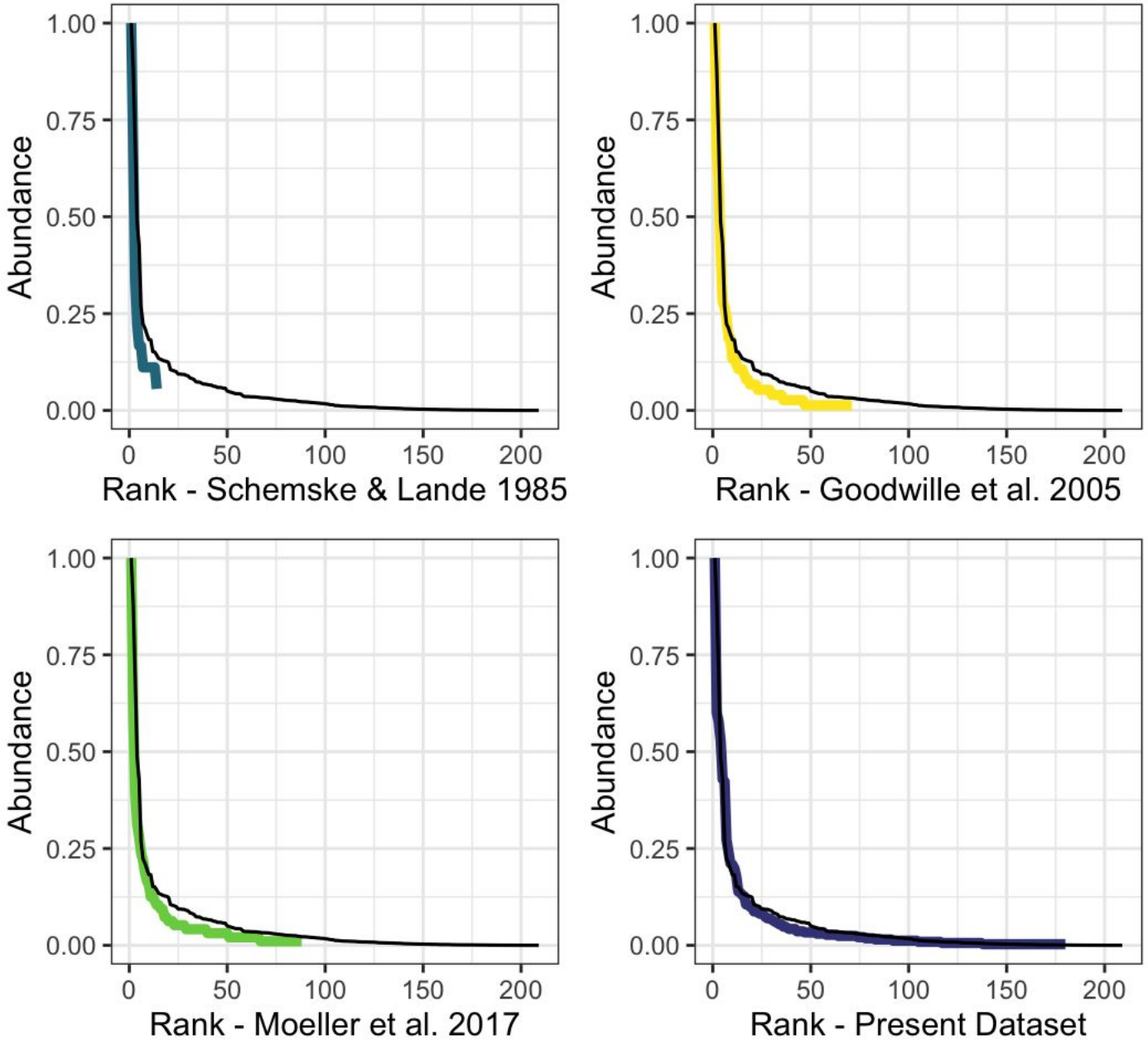

Figure 3. Rank-abundance plots (option 1 – not log scale). Rank-abundance curves are shown for three cumulative studies which build on the same dataset, and the present study.

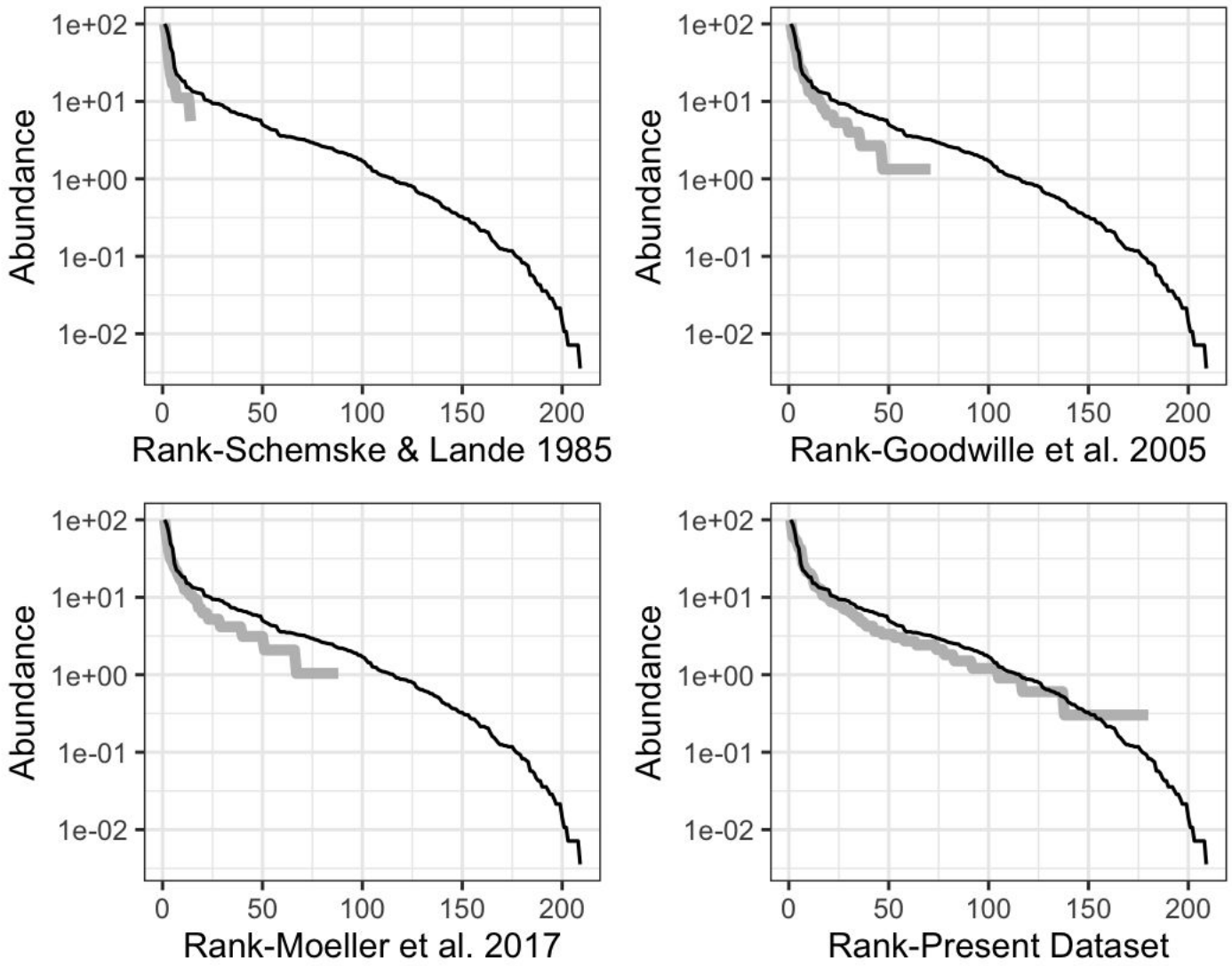

Figure 3. Rank-abundance plots – option 2, log-transformed, non-data color removed.

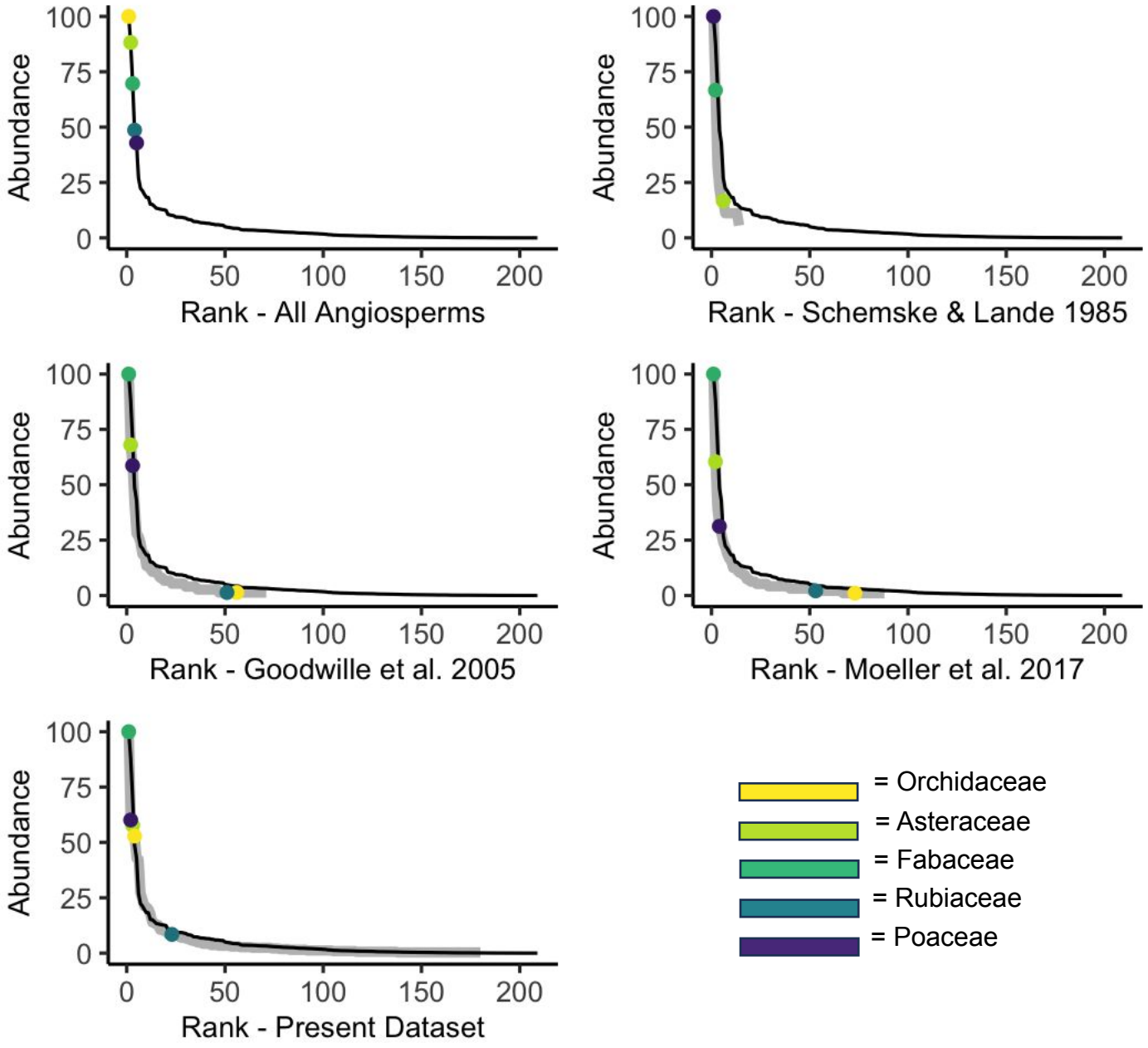

Figure 4. Rank-abundance curves for Schemske and Lande (1985), Goodwille et al. (2005), Moeller et al. (2017), and the present study, showing the relative positions of the most abundant angiosperm families in their respective datasets. The top left panel shows the expected abundance based on relative angiosperm family diversity. (Note that Orchidaceae and Rubiaceae do not appear in Schemske and Lande (1985) and thus are not plotted on the curve.)

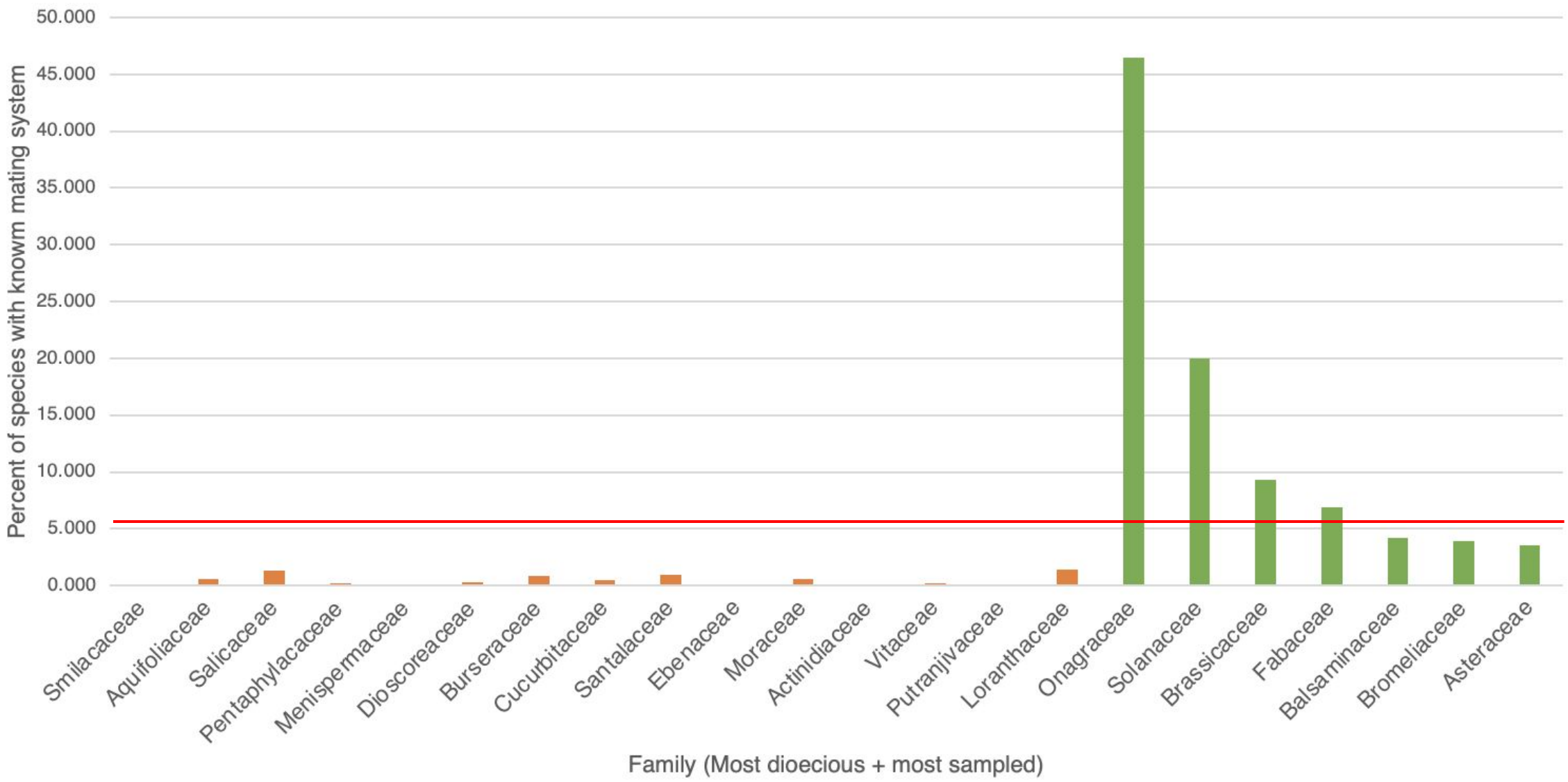

Figure 5: The percent coverage for these families is shown alongside six of the best-covered families for scale. The red line represents the average percent coverage across all families, at 6.95% coverage.

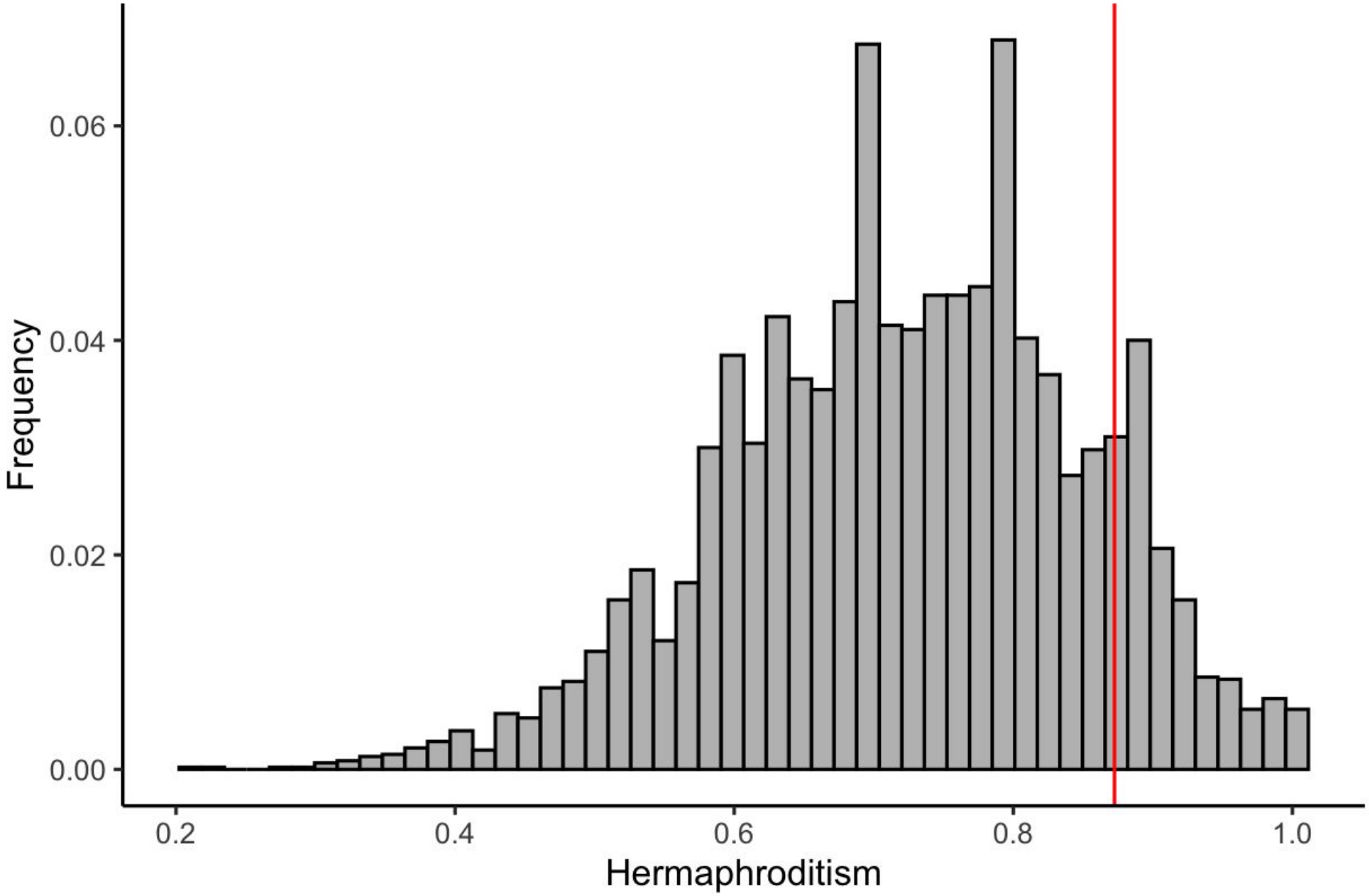

Figure 6.

Red line = mean for overrepresented families, = 0.8725