

# Social Network Analysis of the Genealogy of Strawberry: Retracing the Wild Roots of Heirloom and Modern Cultivars

Dominique D.A. Pincot<sup>\*,1</sup>, Mirko Ledda<sup>\*,1</sup>, Mitchell J. Feldmann<sup>\*,1</sup>, Michael A. Hardigan<sup>\*</sup>, Thomas J. Poorten<sup>\*</sup>, Daniel E. Runcie<sup>\*</sup>, Christopher Heffelfinger<sup>†</sup>, Stephen L. Dellaporta<sup>†</sup>, Glenn S. Cole<sup>\*</sup> and Steven J. Knapp<sup>\*,2</sup>

<sup>\*</sup>Department of Plant Sciences, University of California, Davis, One Shields Avenue, Davis, California, 95616, USA, <sup>†</sup>Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, Connecticut, 06520, USA

**ABSTRACT** The widely recounted story of the origin of cultivated strawberry (*Fragaria* × *ananassa*) oversimplifies the complex interspecific hybrid ancestry of the highly admixed populations from which heirloom and modern cultivars have emerged. To develop deeper insights into the three century long domestication history of strawberry, we reconstructed the genealogy as deeply as possible—pedigree records were assembled for 8,851 individuals, including 2,656 cultivars developed since 1775. The parents of individuals with unverified or missing pedigree records were accurately identified by applying exclusion analysis to array-genotyped single nucleotide polymorphisms. We identified 187 wild octoploid and 1,171 *F.* × *ananassa* founders in the genealogy, from the earliest hybrids to modern cultivars. The pedigree networks for cultivated strawberry are exceedingly complex labyrinths of ancestral interconnections formed by diverse hybrid ancestry, directional selection, migration, admixture, bottlenecks, overlapping generations, and recurrent hybridization with common ancestors that have unequally contributed allelic diversity to heirloom and modern cultivars. Fifteen to 333 ancestors were predicted to have transmitted 90% of the alleles found in country-, region-, and continent-specific populations. Using parent-offspring edges in the global pedigree network, we found that selection cycle lengths over the last 200 years of breeding have been extraordinarily long (16.0-16.9 years/generation) but decreased to a present-day range of 6.0-10.0 years/generation. Our analyses uncovered conspicuous differences in the ancestry and structure of North American and European populations and shed light on forces that have shaped phenotypic diversity in *F.* × *ananassa*.

**KEYWORDS** *Fragaria*; kinship; domestication; DNA forensics; biodiversity; conservation genetics

The strawberries found in markets around the world today are produced by cultivated strawberry (*Fragaria* × *ananassa* (Weston) Duchesne ex Rozier), a species domesticated over the last 300 years (Darrow 1966). *F.* × *ananassa* is technically not a species but an admixed population of interspecific hybrid lineages between cross-compatible wild allo-octoploid ( $2n = 8x = 56$ ) species with shared evolutionary histories (Duchesne 1766; Darrow 1966; Liston *et al.* 2014). The earliest *F.* × *ananassa* cultivars originated as spontaneous hybrids between *F. chiloensis*

and *F. virginiana* in Brittany, the Garden of Versailles, and other Western European gardens in the early 1700s, shortly after the migration of *F. chiloensis* from Chile to France in 1714 (Duchesne 1766; Bunyard 1917; Darrow 1966; Pitrat and Fauray 2003). Their serendipitous origin was discovered by the French botanist Antoine Nicolas Duchesne (1747-1827) and famously described in a treatise on strawberries that biologists suspect included one of the first renditions of a phylogenetic tree (Duchesne 1766). Even though those studies pre-dated both the advent of genetics and the discovery of ploidy differences in the genus, the phylogenies were remarkably close to hypotheses that emerged more than 150 years later (Darrow 1966; Staudt 1988, 2003; Dillenberger *et al.* 2018). The early interspecific hybrids were observed to be phenotypically unique and horticulturally superior to their

doi: 10.1534/genetics.XXX.XXXXXX

Manuscript compiled: Wednesday 30<sup>th</sup> September, 2020

<sup>1</sup>These authors contributed equally to this work.

<sup>2</sup>Corresponding author: Department of Plant Sciences, University of California, Davis, One Shields Avenue, Davis, California, 95616, USA; sjknapp@ucdavis.edu.

1 wild octoploid parents, which drove the domestication of *F.* ×  
2 *ananassa*. [Hardigan et al. \(2020a,b\)](#) showed that hybrids between  
3 *F. chiloensis* and *F. virginiana* had nearly double the heterozygos-  
4 ity of their parents, which almost certainly boosted phenotypic  
5 variation and fueled *F.* × *ananassa* domestication. The cultivation  
6 of *F.* × *ananassa* steadily increased and ultimately supplanted  
7 the cultivation of other strawberry species, forever changing  
8 strawberry production and consumption worldwide ([Fletcher](#)  
9 [1917](#); [Darrow 1966](#); [Wilhelm and Sagen 1974](#); [Finn et al. 2013](#)).

10 The romanticized and widely recounted story of the origin  
11 of cultivated strawberry, while compelling, oversimplifies the  
12 complexity of the wild ancestry and 300-year history of do-  
13 mestication ([Darrow 1966](#)). The domestication of *F.* × *ananassa*  
14 has been documented in narrative histories and pedigree- and  
15 genome-informed studies of genetic diversity and population  
16 structure, but has not been fully untangled or deeply studied  
17 ([Clausen 1915](#); [Fletcher 1917](#); [Darrow 1966](#); [Wilhelm and Sagen](#)  
18 [1974](#); [Sjulin and Dale 1987](#); [Bringhurst et al. 1990](#); [Dale and Sjulin](#)  
19 [1990](#); [Johnson 1990](#); [Sjulin 2006](#); [Hancock et al. 2008](#); [Horvath et al.](#)  
20 [2011](#); [Sánchez-Sevilla et al. 2015](#)). The only pedigree-informed  
21 studies of the breeding history of cultivated strawberry focused  
22 on an analysis of the ancestry of 134 North American cultivars  
23 developed between 1960 and 1985 ([Sjulin and Dale 1987](#); [Dale](#)  
24 [and Sjulin 1990](#)). They identified 53 founders in the pedigrees  
25 of those cultivars and estimated that 20 founders contributed  
26 approximately 85% of the allelic diversity. The inference reached  
27 in those studies and others was that cultivated strawberry is  
28 genetically narrow ([Sjulin and Dale 1987](#); [Dale and Sjulin 1990](#);  
29 [Hancock and Luby 1995](#); [Graham et al. 1996](#); [Hancock et al. 2001](#);  
30 [Hummer 2008](#); [Gaston et al. 2020](#)). The genetic narrowness hy-  
31 pothesis, however, has not been supported by genome-wide  
32 analyses of DNA variants, which have shown that *F. chiloensis*,  
33 *F. virginiana*, and *F.* × *ananassa* harbor massive nucleotide diver-  
34 sity and that a preponderance of the alleles transmitted by the  
35 wild octoploid founders have survived domestication and been  
36 preserved in the global *F.* × *ananassa* population ([Hardigan et al.](#)  
37 [2020a,b](#)).

38 The domestication of cultivated strawberry has followed a  
39 path quite different from that of other horticulturally impor-  
40 tant species, many of which were domesticated over millen-  
41 nia and trace to early civilizations, e.g., apple (*Malus domes-*  
42 *tica*), olive (*Olea europaea* subsp. *europaea*), and wine grape (*Vitis*  
43 *vinifera* subsp. *vinifera*) ([Purugganan and Fuller 2009](#); [Myles et al.](#)  
44 [2011](#); [Meyer et al. 2012](#); [Meyer and Purugganan 2013](#); [Cornille](#)  
45 [et al. 2014](#); [Larson et al. 2014](#); [Diez et al. 2015](#); [Duan et al. 2017](#)).  
46 Although the octoploid progenitors were cultivated before the  
47 emergence of *F.* × *ananassa*, the full extent of their cultivation  
48 is unclear and neither appears to have been intensely domesti-  
49 cated, e.g., [Hardigan et al. \(2020b\)](#) did not observe changes in  
50 the genetic structure between land races and wild ecotypes of  
51 *F. chiloensis*, a species cultivated in Chile for at least 1,000 years  
52 ([Finn et al. 2013](#)). With less than 300 years of breeding, pedigrees  
53 for thousands of *F.* × *ananassa* individuals have been recorded,  
54 albeit in disparate sources. To delve more deeply into the domes-  
55 tication history of cultivated strawberry, we assembled pedigree  
56 records from hundreds of sources and reconstructed the geneal-  
57 ogy as deeply as possible. One of the original impetuses for  
58 this study was to identify historically important and genet-  
59 ically prominent ancestors for whole-genome shotgun (WGS)  
60 resequencing and genome-scale analyses of nucleotide diversity  
61 ([Hardigan et al. 2020a,b](#)).

62 One challenge we faced when building the pedigree database

and reconstructing the genealogy of strawberry was the ab-  
63 sence of pedigree records for 96% of the 1,287 accessions pre-  
64 served in the University of California, Davis (UCD) Strawberry  
65 Germplasm Collection, hereafter identified as the ‘California’  
66 population. To solve this problem, authenticate pedigrees, and  
67 reconstruct the genealogy of the California population, we ap-  
68 plied exclusion analysis in combination with high-density single  
69 nucleotide polymorphism (SNP) genotyping ([Chakraborty](#)  
70 [et al. 1974](#); [Elston 1986](#); [Goldgar and Thompson 1988](#); [Pena and](#)  
71 [Chakraborty 1994](#); [Vandeputte 2012](#); [Vandeputte and Haffray](#)  
72 [2014](#)). Here, we describe the accuracy of parent identification  
73 by exclusion analysis among individuals genotyped with 35K,  
74 50K, or 850K SNP arrays ([Bassil et al. 2015](#); [Verma et al. 2016](#);  
75 [Hardigan et al. 2020a](#)). Several thousand SNP markers common  
76 to the three arrays were integrated to develop a SNP profile  
77 database for the exclusion analyses described here.

78 The genealogies (pedigree networks) of domesticated plants,  
79 especially those with long-lived individuals, overlapping gener-  
80 ations, and extensive migration and admixture, can be challeng-  
81 ing to visualize and comprehend ([Mäkinen et al. 2005](#); [Trager](#)  
82 [et al. 2007](#); [Voorrips et al. 2012](#); [Shaw et al. 2014](#); [Fradgley et al.](#)  
83 [2019](#); [Muranty et al. 2020](#)). We used Helium ([Shaw et al. 2014](#))  
84 to visualize certain targeted pedigrees; however, the strawberry  
85 pedigree network was too large and complex to be effectively  
86 visualized and analyzed with traditional pedigree visualization  
87 approaches.

88 The pedigree networks of plants and animals share many  
89 of the features of social networks with nodes (individuals) con-  
90 nected to one another through edges (parent-offspring relation-  
91 ships) ([Barabási et al. 2011](#); [Barabási 2016](#); [Contandriopoulos et al.](#)  
92 [2018](#)). We used social network analysis (SNA) methods, in com-  
93 bination with classic population genetic methods, to analyze  
94 the genealogy and develop deeper insights into the domestica-  
95 tion history of strawberry ([Lacy 1989, 1995](#); [Barabási et al. 2011](#);  
96 [Barabási 2016](#); [Contandriopoulos et al. 2018](#)). SNA approaches  
97 have been applied in diverse fields of study but have apparently  
98 not yet been applied to the problem of analyzing and charac-  
99 terizing pedigree networks ([Moreno 1953](#); [Scott 1988](#); [Edwards](#)  
100 [1992](#); [Wasserman and Faust 1994](#); [Kominakis 2001](#)). With SNA,  
101 narrative data (birth certificates and pedigree records) are trans-  
102 lated into relational data (parent-offspring and other genetic  
103 relationships) and summary statistics (betweenness centrality  
104 and out-degree) and visualized as sociograms (pedigree net-  
105 works) ([Barabási et al. 2011](#); [Barabási 2016](#); [Contandriopoulos](#)  
106 [et al. 2018](#)). Here, we report insights gained from studies of  
107 the formation and structure of domesticated populations world-  
108 wide, the complex wild ancestry of *F.* × *ananassa*, and genetic  
109 relationships among extinct and extant ancestors in demograph-  
110 ically unique domesticated populations tracing to the earliest  
111 hybrids ([Darrow 1966](#)).

## 113 Materials and Methods

### 114 Pedigree Record Assembly, Documentation, and Annotation

115 We located and assembled pedigree records for strawberry ac-  
116 cessions from more than 807 documents, databases, and other  
117 sources including: (a) US Patent and Trademark Office Plant  
118 Patents (<https://www.uspto.gov/>); (b) Germplasm Resource and  
119 Information Network (GRIN) passport data for accessions pre-  
120 served in the USDA National Plant Germplasm System (NPGS;  
121 <https://www.ars-grin.gov/>); (c) the original unpublished UCD lab-  
122 oratory notebooks and other documents of Royce S. Bringhurst  
123 archived in a special collection at the Merrill-Cazier Library,

1 Utah State University, Logan, Utah (Bringhurst 1918-2016; USU  
2 COLL MSS 515; [http://archiveswest.orbiscascade.org/ark:/80444/  
3 xv47241](http://archiveswest.orbiscascade.org/ark:/80444/xv47241)); (d) the original unpublished University of Califor-  
4 nia, Berkeley (UCB) laboratory notebooks of Harold E. Thomas  
5 loaned by Phillip Stewart (Driscoll's, Watsonville, California);  
6 (e) an obsolete electronic database discovered and recovered  
7 at UCD; (f) an electronic pedigree database for public cultivars  
8 developed by Thomas Sjulín, a former strawberry breeder at  
9 Driscoll's, Watsonville, California; (g) scientific, technical bul-  
10 letins, and popular press articles; and (h) garden catalogs (Files  
11 S1-S3).

12 The pedigree records and other input data were manually  
13 curated and deduplicated. The database was constructed in a  
14 standard trio format (offspring, mother, father) with supporting  
15 passport data, which included: (a) alphanumeric identification  
16 numbers; (b) common names or aliases; (c) accession types (e.g.,  
17 cultivars, breeding materials, or wild ecotypes); (d) birth years  
18 (years of origin); (e) geographic origin; (f) inventor (breeder or  
19 institution) names; (g) taxonomic classifications, and (h) DNA-  
20 authenticated pedigrees for genotyped UCD accessions, as de-  
21 scribed below (File S1). Because a parent could be a male in one  
22 cross and female in another, and parent sexes were frequently  
23 unknown or inconsistently recorded in pedigree records, the  
24 'mother' (parent 1) and 'father' (parent 2) designations were  
25 arbitrary and unimportant to our study.

26 Germplasm accession numbers in the pedigree database in-  
27 cluded 'plant introduction' (PI) numbers for USDA accessions,  
28 UCD identification numbers for UCD accessions, and assorted  
29 other identification numbers. UCD accession numbers were  
30 written in a 10-digit machine-readable and searchable format  
31 to convey birth year and unique numbers, e.g., the UCD ID  
32 '65C065P001' identifies a single individual (P001) in full-sib fam-  
33 ily C065 born in 1965 that was identified in historic records  
34 as '65.65-1' (Bringhurst 1918-2016; Bringhurst *et al.* 1980). The  
35 latter is the 'Bringhurst' notation found in the historic pedi-  
36 gree records for UCD accessions and US Plant Patents. The  
37 decimals and dashes in the original notation created problems  
38 with data curation, analysis, and sorting. To solve this, the  
39 original 'Bringhurst' accession numbers (e.g., 65.65-1) were con-  
40 verted into the 10-digit machine-readable accession numbers  
41 (e.g., 65C065P001) reported in our pedigree database, where 'C'  
42 identifies a cultivated strawberry accession. Common names  
43 (aliases) of cultivars and accessions (if available) were concate-  
44 nated with underscores to create machine-readable and sortable  
45 names, e.g., the name for the *F. × ananassa* cultivar 'Madame  
46 Moutot' was stored as 'Madame\_Moutot'. Cultivars sharing  
47 names were made unique by appending an underscore and their  
48 year. Throughout the pedigree database, unknown individuals  
49 were created as necessary and identified with unique alphanu-  
50 meric identification numbers starting with the prefix 'Unknown',  
51 followed by an underscore, a species acronym when known or  
52 NA when unknown, an underscore, and consecutive numbers,  
53 e.g., 'Unknown\_FC\_071' identifies unknown *F. chiloensis* founder  
54 71. The species acronyms applied in our database were FA for *F.*  
55 *× ananassa*, FC for *F. chiloensis*, FV for *F. virginiana*, FW for *F. vesca*  
56 (woodland strawberry), FI for *F. iinumae*, FN for *F. nipponica*, FG  
57 for *F. viridis* (green strawberry), FM for *F. moschata*, and FX for  
58 other wild species or interspecific hybrids, e.g., *F. × vescana*.

### 59 Plant Material and SNP Profile Database

60 To develop a SNP profile database for DNA forensic and popu-  
61 lation genetic analyses (see below), we recalled and reanalyzed

62 SNP marker genotypes for 1,495 individuals, including 1,235  
63 UCD and 260 USDA accessions (asexually propagated individ-  
64 uals) previously genotyped by Hardigan *et al.* (2018) with the  
65 iStraw35 SNP array (Bassil *et al.* 2015; Verma *et al.* 2016). SNP  
66 marker genotypes were automatically called with the Affymetrix  
67 Axiom Analysis Suite (v1.1.1.66, Affymetrix, Santa Clara, CA).  
68 DNA samples with > 6% missing data were dropped from our  
69 analyses. We used quality metrics output by the Affymetrix  
70 Axiom Analysis Suite and custom R scripts and the R pack-  
71 age *SNPRelate* (Zheng *et al.* 2012) to identify and select codomi-  
72 nant SNP markers with genotypic clustering confidence scores  
73  $(1 - p_C) \geq 0.01$ , where  $p_C$  is the posterior probability that the  
74 SNP genotype for an individual was assigned to the correct geno-  
75 typic cluster (Affymetrix Inc. 2015). This yielded 14,650 high  
76 confidence co-dominant SNP markers for paternity-maternity  
77 analyses. While SNP markers are co-dominant by definition,  
78 a certain percentage of the SNP markers assayed in a popula-  
79 tion produce genotypic clusters lacking one of the homozygous  
80 genotypic clusters. These so-called 'no minor homozygote' SNP  
81 markers were excluded from our analyses.

82 For a second DNA forensic analysis, 1,561 UCD individuals  
83 were genotyped with 50K or 850K SNP arrays (Hardigan *et al.*  
84 2020a). This study population included 560 hybrid offspring  
85 from crosses among 27 elite UCD parents, the *F. × ananassa* culti-  
86 var 'Puget Reliance', and the *F. chiloensis* subsp. *lucida* ecotypes  
87 'Del Norte' and 'Oso Flaco'. Hardigan *et al.* (2020a) included  
88 16,554 SNP markers from the iStraw35 and iStraw90 SNP arrays  
89 on the 850K SNP array. To build a SNP profile database for the  
90 second paternity-maternity analysis, we identified 2,615 SNP  
91 markers that were common to the three arrays and produced  
92 well separated co-dominant genotypic clusters with high con-  
93 fidence scores ( $p_C > 0.99$ ) and < 6% missing data (Bassil *et al.*  
94 2015; Verma *et al.* 2016; Hardigan *et al.* 2020a).

95 We subdivided the global population (entire pedigree) into  
96 'California' and 'cosmopolitan' populations, in addition to  
97 continent-, region-, or country-specific populations, for different  
98 statistical analyses. These subdivisions are documented in the  
99 pedigree database (File S1). The California population included  
100 100% of the UCD individuals ( $n = 3,540$ ) from the global popu-  
101 lation, in addition to 262 non-California individuals that were  
102 ascendants of UCD individuals. The cosmopolitan population  
103 included 100% of the non-California (non-UCD) individuals  
104 ( $n = 5,193$ ), in addition to 160 California individuals that were  
105 ascendants of non-California individuals. We subdivided indi-  
106 viduals in the US population (excluding UCD individuals) into  
107 Midwestern, Northeastern, Southern, and Western US popula-  
108 tions. The Western US population included only those UCD  
109 individuals that were ascendants in the pedigrees of Western US  
110 individuals. The country specific subdivisions were Australia,  
111 China, Japan, South Korea, Belgium, Czechoslovakia, Denmark,  
112 England, Finland, France, Germany, Israel, Italy, the Nether-  
113 lands, Norway, Poland, Russia, Scotland, Spain, Sweden, and  
114 Canada.

### 115 DNA Forensic Analyses

We applied standard DNA forensic approaches for diploid organ-  
isms to the problem of identifying parents and authenticating  
pedigrees in allo-octoploid strawberry (Chakraborty *et al.* 1974;  
Elston 1986; Jones and Ardren 2003; Telfer *et al.* 2015; Muranty  
*et al.* 2020). Genotypic transgression ratios were estimated for all  
possible duos and trios of individuals in two study populations  
(described above) from genotypes of multiple SNP marker loci.

For duos of individuals in the SNP profile database for a population, the genotypic transgression score for the  $i$ th SNP marker was estimated by

$$S_i = f(AA_{O_i}) \cdot f(BB_{P_i}) + f(BB_{O_i}) \cdot f(AA_{P_i}) \quad (1)$$

where  $i = 1, 2, \dots, m$ ,  $m$  = number of SNP marker loci genotyped in each pair of probative DNA samples,  $f(--_{O_i})$  is the frequency of a homozygous genotype (coded  $AA$  and  $BB$ ) in the candidate offspring individual and  $f(--_{P_i})$  is the frequency of a homozygous genotype in the candidate parent individual (similarly coded  $AA$  and  $BB$ ) for the  $i$ th SNP marker locus. This equation was applied to a single pair of candidate individuals at a time and was thus constrained to equal 0 or 1; hence,  $S_i = 0$  when homozygous genotypes were identical for a pair of individuals and  $S_i = 1$  when homozygous genotypes were different for a pair of individuals. Duo-transgression ratios (*DTRs*) were estimated for every pair of individuals in the population by summing  $S_i$  estimates from equation (1) over  $m$  marker loci:

$$DTR = \frac{1}{m} \sum_{i=1}^m S_i \quad (2)$$

For trios of individuals in the SNP profile database for a population, the genotypic transgression score for the  $i$ th SNP marker was estimated by

$$T_i = f(AB_{O_i}) \cdot f(AA_{P_1}) \cdot f(AA_{P_2}) + f(AB_{O_i}) \cdot f(BB_{P_1}) \cdot f(BB_{P_2}) \quad (3)$$

where  $f(AB_{O_i})$  is the frequency of a heterozygous genotype (coded  $AB$ ) in the candidate offspring individual,  $f(--_{P_1})$  is the frequency of either homozygous genotype ( $AA$  or  $BB$ ) in candidate parent 1 ( $P_1$ ), and  $f(--_{P_2})$  is the frequency of either homozygous genotype in candidate parent 2 ( $P_2$ ) for the  $i$ th SNP marker locus. Trio transgression ratios (*TTRs*) were estimated for every trio of individuals in the population by summing  $T_i$  estimates from equation (3) over  $m$  marker loci:

$$TTR = \frac{1}{m} \sum_{i=1}^m T_i + S1_i + S2_i - S1_i \cdot S2_i \quad (4)$$

where  $m$  is the number of SNP marker loci genotyped for a trio of individuals,  $S1_i$  is the score estimated from equation (1) for candidate parent 1, and  $S2_i$  is the score estimated from equation (1) for candidate parent 2. To avoid double counting transgressions, *TTR* estimates were corrected by subtracting  $S1_i \times S2_i$ .

Our analyses yielded *DTR* and *TTR* estimates for paternity and maternity exclusion tests among genotyped individuals in the study populations. The putative parents of offspring were identified by estimating the probability of paternity (or maternity) from equations (2) and (4) and empirically estimating statistical significance thresholds by bootstrapping—50,000 bootstrap samples of size  $n$  were drawn with replacement from  $n$  probative DNA samples of individuals with declared parents in the population (Efron 1982; Simon and Bruce 1991; Manly 2006; Berry *et al.* 2014). The ‘declared’ or ‘stated’ parents are those recorded in pedigree records, whereas the ‘DNA-authenticated’ parents are those verified by exclusion analysis. The bootstrap-estimated *TTR*-threshold of  $TTR \leq 0.01$  yielded false-positive and negative probabilities of zero when estimated by summing  $T_i$  estimates over 14,650 SNP marker loci. Similarly, the bootstrap-estimated *DTR*-threshold of  $DTR \leq 0.0016$  yielded a false positive probability of zero and a false negative probability of 5% when estimated by summing  $S_i$  estimates over 14,650 SNP marker loci.

## Social Network Analyses

The pedigree networks for global, California, and cosmopolitan populations were analyzed and visualized as directed social networks using the R package *igraph* (version 1.2.2; Csardi and Nepusz 2006), where every edge in the graph connects a parent node to an offspring node and information flows unidirectionally from parents to offspring (Wasserman and Faust 1994). The pedigree networks or sociograms were visualized using the open-source software Gephi (version 0.9.2; Bastian *et al.* 2009; <https://gephi.org/>). We estimated the number of edges ( $d$  = degree) and in-degree ( $d_i$ ), out-degree ( $d_o$ ), and betweenness centrality ( $B$ ) statistics for every individual in a sociogram (Wasserman and Faust 1994).  $d_i$  estimates the number of known parents, where  $d_i = 0$  when neither parent is known (for founders), 1 when one parent is known, and 2 when both parents are known.  $d_o$  estimates the number of descendants of an individual. A ‘geodesic’ is the shortest path between two nodes in the network and estimates the number of generations in the pedigree of an individual (Hayes 2000).  $D$  is the longest geodesic in the network and estimates the largest number of generations for a descendant in the pedigree or the maximum depth of the pedigree (Hayes 2000).  $B$  estimates the connectivity of an individual to other individuals in a network (the number of geodesics connecting a node to other nodes), essentially the flow of information (alleles) and information ‘bottlenecks’ (Freeman 1977; Wasserman and Faust 1994; Yu *et al.* 2007; Pavlopoulos *et al.* 2011).  $B$  was estimated by

$$B(n_i) = \sum_{j < k} \frac{g_{jk}(n_i)}{g_{jk}} \quad (5)$$

where  $n_i$  is the  $i$ th node (individual),  $i$ ,  $j$ , and  $k$  are different nodes,  $g_{jk}$  is the number of geodesics occurring between nodes  $j$  and  $k$ , and  $g_{jk}(n_i)$  is the number of geodesics that pass through the  $i$ th node (Freeman 1977; Wasserman and Faust 1994; Brandes 2001; Csardi and Nepusz 2006).  $B = 0$  when  $d_i$  or  $d_o$  equal zero.

Standard social network analysis metrics and terminology were used to classify individuals and describe their importance in the genealogy, which are analogous to applications in diverse fields of study (Gursoy *et al.* 2008; Koschützki and Schreiber 2008; Morselli 2010; Kim and Song 2013; Nerghees *et al.* 2015). Using  $B$  and  $d_o$  estimates, ancestors were classified as globally central ( $d_o > \bar{d}_o \wedge B > \bar{B}$ ), locally central ( $d_o > \bar{d}_o \wedge B < \bar{B}$ ), broker ( $d_o < \bar{d}_o \wedge B > \bar{B}$ ), or marginal ( $d_o < \bar{d}_o \wedge B < \bar{B}$ ).

## Selection Cycle Length Calculations

The pedigree network for every cultivar was extracted from the global pedigree network and included the cultivar (the youngest terminal node) and every ascendant (founder and non-founder) of the cultivar. Selection cycle lengths ( $S$  = years/generation) were estimated for every cultivar by tracing every possible path (back in time) in the pedigree network from the cultivar to founders and calculating birth year differences for every parent-offspring edge ( $y_i$ ) in the path, where  $y_i$  is the number of years separating the  $i$ th parent-offspring edge. The mean selection cycle length was estimated by  $\bar{S} = \sum_i y_i / n_e$ , where  $y_i$  is the birth year difference for the  $i$ th parent-offspring edge,  $n_e$  is the number of parent-offspring edges and  $i = 1, 2, \dots, n_e$ . To understand how selection cycle length changed over time, we considered all 14,275 unique parent-offspring edges available in the pedigree, among which 9,486 had birth years known for both the parent and the offspring. For each edge, we computed its midpoint as the average birth year between the parent and the offspring and

1 its size, i.e. the selection cycle length ( $S$ ), as the difference in  
2 birth years between the parent and the offspring.

### 3 **Estimation of Coancestry and Pedigree-Genomic Relationship** 4 **Matrices**

5 The kinship or coancestry matrix ( $A$ ) was estimated for the entire  
6 pedigree ( $n = 8,851$  individuals) using the *create.pedigree*  
7 and *kin* functions in the R package *synbreed* (version 0.12-12;  
8 [Wimmer et al. 2012](#)), where the  $i$ th diagonal element of  $A$  is the  
9 coefficient of coancestry of individual  $i$  with itself ( $C_{ii}$ ) and the  
10  $ij$ th off-diagonal element of  $A$  is the coefficient of coancestry  
11 between individuals  $i$  and  $j$  ( $C_{ij}$ ) ([Lynch and Walsh 1998](#)). The  
12 genomic relationship matrix ( $G$ ) was estimated for 1,495 individ-  
13 uals genotyped with 14,650 SNP markers selected to have minor  
14 allele frequencies (MAF)  $\geq 0.05$  and  $\leq 10\%$  missing data.  $G$  was  
15 estimated as described by [VanRaden \(2008\)](#) using the function  
16 *A.mat* in the R package *rrBLUP* (version 4.6.1; [Endelman 2011](#)).  
17 Missing genotypes were imputed using the mean genotype for  
18 each SNP marker.

19 We estimated the combined pedigree-genomic relationship  
20 matrix ( $H$ ) for the entire pedigree ( $n = 8,851$  individuals) as  
21 described by [Legarra et al. \(2009\)](#). The  $A$  matrix was partitioned  
22 into four sub-matrices ( $A_{11}, A_{12}, A_{21},$  and  $A_{22}$ ), where the sub-  
23 script 1 indexes ungenotyped and 2 indexes genotyped individ-  
24 uals.  $G$  and  $A_{22}$  had the same dimensions but different scales.  
25 To construct the scaled  $G$  matrix ([Christensen 2012](#); [Christensen](#)  
26 [et al. 2012](#); [Gao et al. 2012](#)), the mean of off-diagonal elements  
27 of  $G$  ( $\overline{oG}$ ) were scaled to match  $\overline{oA_{22}}$  and the mean of diagonal  
28 elements of  $G$  ( $\overline{dG}$ ) were scaled to match  $\overline{dA_{22}}$ :

$$\overline{dG}\beta + \alpha = \overline{dA_{22}}$$

and

$$\overline{oG}\beta + \alpha = \overline{oA_{22}}$$

with scalar solutions

$$\alpha = \overline{oA_{22}} - \overline{oG}\beta$$

and

$$\beta = \frac{\overline{dA_{22}} - \overline{oA_{22}}}{\overline{dG} - \overline{oG}}$$

29 The  $H$  matrix was estimated using the scaled  $G$  matrix ( $\tilde{G} =$   
30  $G\beta + \alpha$ ) as described by [Legarra et al. \(2009\)](#):

$$H = \begin{bmatrix} A_{11} + A_{12}A_{22}^{-1}(\tilde{G} - A_{22})A_{22}^{-1}A_{21} & A_{12}A_{22}^{-1}\tilde{G} \\ \tilde{G}A_{22}^{-1}A_{21} & \tilde{G} \end{bmatrix} \quad (6)$$

31 The open-source R code we developed to estimate  $H$  has been  
32 deposited in a FigShare database (File S6).

33 To study genetic relationships among extinct and extant indi-  
34 viduals, we estimated separate  $H$  matrices for the California  
35 and cosmopolitan populations and applied principal component  
36 analysis (PCA) to the unscaled  $H$  matrices. Principal compo-  
37 nents were estimated by spectral decomposition of  $H$  using  
38 the *eigen* function from base R (version 4.0.0), which yielded  
39 eigenvalues, eigenvectors, and component scores. Scores for the  
40 first two principal components were then plotted using the R  
41 package *ggplot2* ([Wickham 2016](#)).

### **Genetic Contributions of Founders and Ancestors**

Coancestry or kinship ( $A$ ) matrices were estimated for indi-  
viduals within continent-, region-, and country-specific focal  
populations using the *create.pedigree* and *kin* functions in the R  
package *synbreed* (version 0.12-9; [Wimmer et al. 2012](#)). Focal pop-  
ulations consisted of cultivars and their ascendants (ancestors).  
Founders are ancestors with unknown parents, which were as-  
sumed to be unrelated ([Lacy 1989, 1995](#); [Hartl and Clark 2007](#)),  
whereas non-founders are ancestors with known parents. Termi-  
nal nodes in a pedigree network (sociogram) are either founders  
or the youngest descendants. The mean kinship between the  $i$ th  
founder and cultivars in a focal population was estimated by

$$MK_i = \sum_j C_{ij}$$

where  $C_{ij}$  = the kinship coefficient between the  $i$ th founder and  
 $j$ th cultivar in a focal population,  $i = 1, 2, \dots, n$ ,  $j = 1, 2, \dots, k$ ,  $n$   
= the number of founders in the focal population, and  $k$  = the  
number of cultivars in the focal population. ([Lacy 1989, 1995](#);  
[Lynch and Walsh 1998](#); [Hartl and Clark 2007](#)). The proportional  
genetic contribution of the  $i$ th founder to a focal population  
was estimated by  $P_i = MK_i / \sum_i MK_i$ . The number of founder  
equivalents ( $F_e$ ) was estimated by  $F_e = 1 / \sum_i MK_i$ , where  $i \in$   
{founder<sub>1</sub>, founder<sub>2</sub>, ..., founder<sub>n</sub>} ([Lacy 1989, 1995](#)). Founder  
equivalents "are the number of equally contributing founders  
that would be expected to produce the same genetic diversity as  
in the population under study" ([Lacy 1989](#)).

The genetic contributions (GC) of ancestors (founders and  
non-founders) to a focal population were estimated by construct-  
ing a directed distance matrix ( $D$ ) with dimensions identical to  
 $A$  ( $n \times n$ ) such that parents appeared in the matrix before off-  
spring (alleles flow from parents to offspring but not *vice versa*).  
We used the directed distance (the number of parent-offspring  
edges between two accessions) to modify  $A$  so that coancestry  
coefficients were only estimated between ancestors and direct  
path cultivars. The directed distance matrix  $D$  was estimated  
using the *distances* function in the R package *igraph* (version 1.2.5;  
[Csardi and Nepusz 2006](#)), where non-zero distances in the  $D$   
matrix were set equal to one. Coancestry coefficients for ascendants  
with no direct path to a cultivar were set equal to zero by taking  
the Hadamard product to generate the corrected coancestry ma-  
trix  $A^* = A \odot D$ , where element  $C_{ij}$  = the coancestry coefficient  
for individual  $i$  with itself ([Hartl and Clark 2007](#)). To estimate  
GC for each ancestor, we applied an iterative approach that en-  
tailed: (i) computing  $D$ ,  $A$ , and  $A^* = A \odot D$  from the current  
pedigree; (ii) estimating  $MK_i$  for each ancestor; (iii) ranking  $MK_i$   
estimates from largest to smallest; (iv) setting  $GC_i = MK_i$  for the  
ancestor with the largest  $MK_i$  estimate; (v) deleting the ances-  
tor with the largest  $MK_i$  estimate and rebuilding the pedigree;  
and (vi) repeating the previous steps until genetic contributions  
( $GC_i$ ) had been estimated for each ancestor. The proportional  
genetic contribution of the  $i$ th ancestor to a focal population was  
estimated by  $P_i = GC_i / \sum_i GC_i$ .

### **Data Availability**

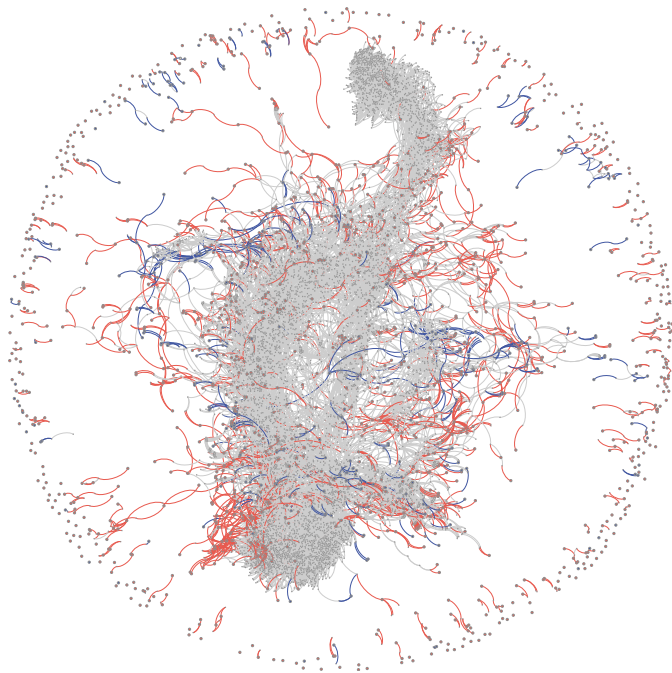
File S1 contains the pedigree database with parents and off-  
spring in a standard trio format (offspring, mother, father) with  
the following passport data: (a) alphanumeric identification  
number; (b) common names or aliases; (c) accession types (e.g.,  
cultivars, breeding materials, or wild ecotypes); (d) birth years  
(years of origin); (e) geographic origins; (f) inventor (breeder or  
institution) names; (g) taxonomic classifications, and (h) DNA-  
authenticated pedigrees for genotyped UCD accessions. File S2

1 contains pedigrees of in the Helium format with parents and  
2 offspring identified by common names or aliases (Shaw *et al.*  
3 2014; <https://github.com/cardinalb/helium-docs/wiki>). File S3 is a  
4 complete bibliography of the databases and documents we refer-  
5 enced to build the pedigree database. Files S4 and S5 contain  
6 betweenness ( $B$ ), in-degree ( $d_i$ ), and out-degree ( $d_o$ ) statistics,  
7 structural role assignments, giant or halo component assign-  
8 ments, and coancestry-based estimates of the genetic contribu-  
9 tions of founders and ancestors to cultivars in the California  
10 and Cosmopolitan populations, respectively. File S6 contains  
11 R code developed to estimate  $H$  from  $A$  and  $G$  as described  
12 by Legarra *et al.* (2009). The example input files from Legarra  
13 *et al.* (2009) for computing the  $H$  matrix are included. File S7  
14 contains R code developed for exclusion (paternity-maternity)  
15 analyses. Table S1 details the most prominent ecotype founders  
16 and their coancestry-based estimates of genetic contribution to  
17 the California and Cosmopolitan populations. All supplements  
18 were uploaded to the FigShare Data Repository.

## 19 Results and Discussion

### 20 Genealogy of Cultivated Strawberry

21 We reconstructed the genealogy of cultivated strawberry as  
22 deeply as possible from wild founders to modern cultivars (Fig.



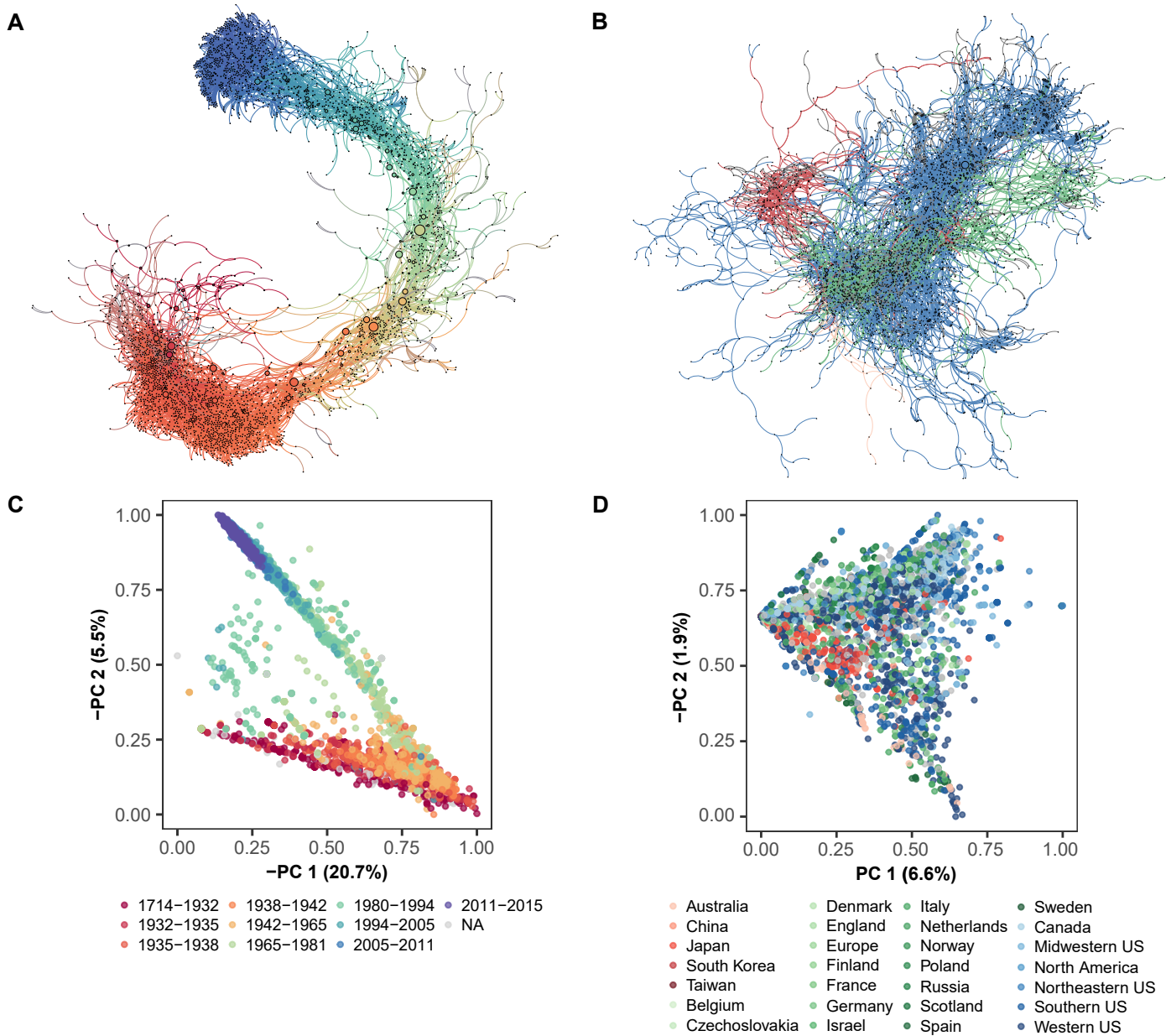
**Figure 1** Global Pedigree Network for Cultivated Strawberry. Sociogram depicting ancestral interconnections among 8,851 accessions, including 8,424 *F. x ananassa* individuals originating as early as 1775, of which 2,656 are cultivars. The genealogy includes *F. chiloensis* and *F. virginiana* founders tracing to 1624 or later. Nodes and edges for 267 wild species founders are shown in blue, whereas nodes and edges for 1,171 *F. x ananassa* founders are shown in red. Founders are individuals with unknown parents. Nodes and edges for descendants (non-founders) are shown in light grey. The outer ring (halo of nodes and edges) are orphans or individuals in short dead-end pedigrees disconnected from the principal pedigree network or so-called 'giant component'.

1; File S1). To build the database, pedigree records for 8,851  
individuals were assembled from more than 800 documents  
including scientific and popular press articles, laboratory note-  
books, garden catalogs, cultivar releases, plant patent databases,  
and germplasm repository databases (Fig. 1; see File S3 for a  
complete bibliography). The database holds pedigree records  
and passport data for 2,656 *F. x ananassa* cultivars, of which  
approximately 310 were private sector cultivars with pedigree  
records in public databases (File S1). The parents of the private  
sector cultivars, however, were nearly always identified by cryp-  
tic alphanumeric codes, and thus could not be integrated into  
the 'giant component' of the sociogram (pedigree network) (Fig.  
1).

The global population was subdivided into 'cosmopolitan'  
and 'California' populations to delve more deeply into their  
unique breeding histories (Hardigan *et al.* 2020b; Fig. 1-2). This  
split was informed by demography and geography, insights  
gained from genome-wide analyses of nucleotide diversity and  
population structure (Hardigan *et al.* 2020a,b), and earlier DNA  
marker-informed studies of genetic diversity (Horvath *et al.* 2011;  
Sánchez-Sevilla *et al.* 2015; Hardigan *et al.* 2018). The cosmopoli-  
tan population included 100% of the non-California (non-UCD)  
individuals ( $n = 5,193$ ) from the global population, in addition  
to 160 California individuals identified as ascendants of non-  
California individuals. The non-California cultivar 'Cascade'  
(PI551759), for example, is a descendant of a cross between the  
California cultivar 'Shasta' (PI551663) and non-California cul-  
tivar 'Northwest' (PI551499) (<https://www.ars.usda.gov/>); hence,  
'Shasta' was included in both the cosmopolitan and California  
populations. Similarly, the California population included 100%  
of the UCD individuals ( $n = 3,540$ ) from the global population,  
in addition to 262 non-California individuals that were identified  
as ascendants of UCD individuals. We nearly completely recon-  
structed the genealogy of the California population; however, as  
described below, pedigree records were missing for nearly every  
individual in the California population but were accurately  
ascertained using computer and DNA forensic approaches.

### 60 Social Network Analyses Uncover Distinctive Differences in 61 the Domestication History of California and Cosmopolitan 62 Populations

We estimated that 80-90% of the individuals in the Califor-  
nia and cosmopolitan pedigree networks were extinct (Fig. 2).  
Using SNP-array genotyped individuals preserved in public  
germplasm collections as anchor points, we searched for evi-  
dence that the allelic diversity transmitted by extinct founders  
had been 'lost'. This is a difficult question to answer with cer-  
tainty; however, the findings reported here, combined with the  
findings of Hardigan *et al.* (2020b), suggest that genetic diversity  
has been exceptionally well preserved in domesticated popula-  
tions. Using SNA and principal component analyses (PCAs)  
of  $H$ , we did not observe structural features in sociograms or  
PCA plots that were indicative of the loss of novel ancestral  
genetic diversity (Fig 2). The kinship or numerator relationship  
matrix ( $A$ ) was estimated for the entire pedigree of genotyped  
and ungenotyped individuals (VanRaden 2008; Legarra *et al.*  
2009). For the present study, 1,495 historically important and  
geographically diverse UCD and USDA *F. x ananassa* individuals  
were genotyped with high-density SNP arrays (Bassil *et al.* 2015;  
Verma *et al.* 2016; Hardigan *et al.* 2020a). The genomic relation-  
ship matrix ( $G$ ) was estimated for the genotyped individuals  
and combined with the  $A$  matrix to estimate the  $H$  matrix for



**Figure 2** Genealogy for California and Cosmopolitan Populations of Cultivated Strawberry. (A) Sociogram depicting ancestral interconnections among 3,802 individuals in the 'California' population. This population included 3,452 *F. × ananassa* individuals developed at the University of California, Davis (UCD) from 1924 to 2012, in addition to 151 non-UCD *F. × ananassa* ascendants that originated between 1775 and 1924. Node and edge colors depict the year of origin of the individual in the pedigree network from oldest (red) to youngest (blue) with a continuous progression from warm to cool colors as a function of time (year of origin). Nodes and edges for individuals with unknown years of origin are shown in grey. (B) Sociogram depicting ancestral interconnections among 5,354 individuals in the 'cosmopolitan' population. This population included 5,106 *F. × ananassa* individuals developed across the globe between 1775 and 2018 and excludes UCD individuals other than UCD ancestors in the pedigrees of non-UCD individuals. Node and edge colors depict the continent where individuals in the pedigree network originated: Australia (orange), Asia (red), North America (blue), and Europe (green). Nodes and edges for individuals of unknown origin are shown in grey. (A and B) For both sociograms, node diameters are proportional to the betweenness centrality (*B*) metrics for individuals (nodes). Orphans and short dead-end pedigrees that were disconnected from the principal pedigree network ('giant component') are not shown. (C) Principal component analysis (PCA) of the pedigree-genomic relationship matrix (*H*) for the California population. The *H* matrix ( $8,851 \times 8,851$ ) was estimated from the coancestry matrix (*A*) for 8,851 individuals and the genomic relationship matrix (*G*) for 1,495 individuals genotyped with a 35K SNP array. The PCA plot shows PC1 and PC2 coordinates for 3,802 individuals in the California population color coded by year-of-origin. (D) PCA of the *H* matrix for the cosmopolitan population. The PCA plot shows PC1 and PC2 coordinates for 5,354 individuals in the cosmopolitan population color coded by country, region, or continent of origin.

1 the entire pedigree (Legarra *et al.* 2009). The global  $H$  matrix  
2 was partitioned as needed for subsequent analyses (Fig. 2).

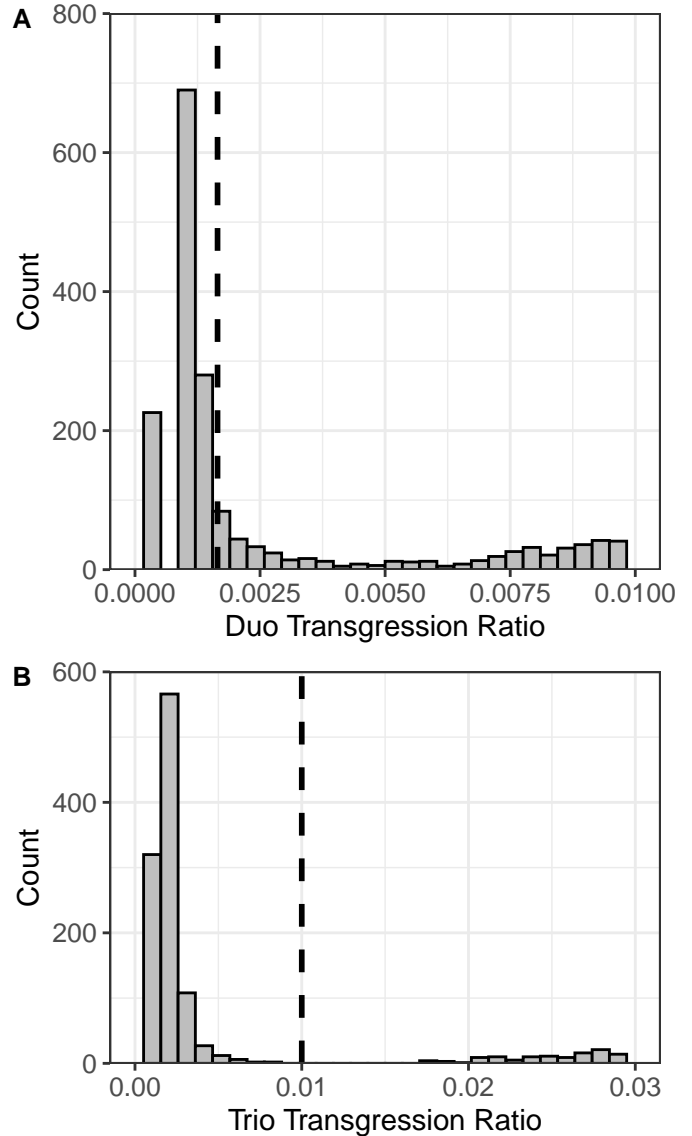
3 PCAs of the  $H$  matrices yielded two-dimensional visualiza-  
4 tions of genetic relationships that were remarkably similar in  
5 shape and structure to sociograms for the California and cos-  
6 mopolitan populations (Fig 2). We observed distinctive differ-  
7 ences in the shapes and structures of the sociograms and PCA  
8 plots between the populations (Fig 2). The pattern in the cos-  
9 mopolitan population was characteristic of pervasive admixture  
10 among individuals across geographies (Fig 2 B and D). We ob-  
11 served a strong chronological trend in the California population  
12 (Fig 2A and C) but not in cosmopolitan population (Fig 2 B  
13 and D). We observed a mid-twentieth century bottleneck in the  
14 California population (the sharp interior angle in the V-shaped  
15 structure of the PCA plot), in addition to a bottleneck pinpointed  
16 to approximately 1987-1993 when the California population be-  
17 came closed. We discovered that 48 founders contributed 100%  
18 of the allelic diversity to the California population from 1987  
19 onward (Fig 2A and C; S1 File). Hardigan *et al.* (2020b) showed  
20 that even though nucleotide diversity had been progressively  
21 reduced by bottlenecks and selection, significant nucleotide di-  
22 versity has persisted in the California population but was found  
23 to be unevenly distributed across the genome.

#### 24 DNA Forensic Approaches for Parent Identification and Pedi- 25 gree Authentication in Octoploid Strawberry

26 When this study was initiated in early 2015, 1,235 *F. × ananassa*  
27 germplasm accessions (asexually propagated individuals) were  
28 preserved in the UCD Strawberry Germplasm Collection. The  
29 collection included 68 UCD cultivars with known pedigrees;  
30 however, pedigree records for the other 1,184 UCD individ-  
31 uals were unavailable. Using computer forensic approaches,  
32 pedigree records for 1,002 individuals were recovered from an  
33 obsolete electronic database. Because the authenticity and ac-  
34 curacy of those records were uncertain, every individual was  
35 genotyped with the iStraw35 SNP array to build a SNP profile  
36 database for parent identification by exclusion analysis (Jones  
37 and Ardren 2003; Vandeputte 2012; Vandeputte and Haffray  
38 2014; Bassil *et al.* 2015; Verma *et al.* 2016). SNP marker genotypes  
39 were automatically called using the Affymetrix Axiom Suite,  
40 then manually curated to identify and extract codominant SNP  
41 markers with well separated genotypic clusters. This yielded  
42 14,650 SNP markers for exclusion analyses. Genotyping errors  
43 were negligible (0.06-0.37%) and genotype-matching percent-  
44 ages for array-genotyped SNPs ranged from 99.63 to 99.95%  
45 among biological and technical replicates.

46 We estimated duo transgression ratios ( $DTRs$ ) for all possi-  
47 ble pairs or duos (761,995) of individuals (Fig. 3). Trio trans-  
48 gression ratios ( $TTRs$ ) were estimated for all possible triplets  
49 or trios of individuals with  $DTR$  estimates in the 0.00 to 0.01  
50 range—individuals with  $DTR$  estimates  $> 0.0016$  were excluded  
51 as parents (Fig. 3). For trio analyses, we included the possibil-  
52 ity that offspring could arise by self-pollination, which yielded  
53  $n \times (n - 1) = 1,235 \times 1,234 = 1,523,990$  possible trios. Al-  
54 though this possibility does not arise in human or animal parent  
55 identification problems (Jones and Ardren 2003; Vandeputte  
56 2012), offspring can arise from self-pollination in cultivated  
57 strawberry and other self-compatible plant species. The number  
58 of possible trios arising from crosses between two parents in  
59 the reference population was  $(n \times [n - 1]) + (n \times [n - 1] \times [n -$   
60  $2])/2 = 941,063,825$ .

61 Trio exclusion analysis identified the parents of 1,044 UCD



**Figure 3** Exclusion Analyses. (A) Distribution of 2,708 duo transgression ratio ( $DTR$ ) estimates falling in the 0.0 to 0.01 range. There were 761,995 possible duos among 1,235 individuals in the California population ( $DTR$  estimates  $> 0.01$  are not shown). The vertical dashed line demarcates the bootstrap-estimated significance threshold ( $DTR < 0.0016$ ) chosen to minimize false positives and negatives. (B) Distribution of 2,815 trio transgression ratio ( $TTR$ ) estimates falling in the 0.00 to 0.03 range. There were 941,063,825 possible  $TTR$  estimates for trios among 1,235 individuals in the California population, which included  $1,235 \times 1,234 = 1,523,990$  possible trios for offspring arising from self-pollination ( $TTR$  estimates  $> 0.03$  are not shown). The vertical dashed line demarcates the bootstrap-estimated significance threshold ( $TTR < 0.01$ ) chosen to minimize false positives and negatives. (A) and (B)  $DTRs$  and  $TTRs$  were estimated by summing over 14,650 SNP markers. Statistical significant thresholds for parent inclusion were empirically estimated from 50,000 bootstrap samples.



1 individuals with 100% accuracy and zero false positives—SNP  
2 profiles for both parents were present in the database for these  
3 individuals (Fig. 3). *DTR* estimates for parents with statistically  
4 significant *TTR* estimates ( $TTR < 0.01$ ) were statistically signifi-  
5 cant ( $DTR < 0.0016$ ). When the SNP profile for only one parent  
6 was present in the database (134 out of 1,235 individuals), duo  
7 exclusion analysis identified those parents with 95% accuracy  
8 and zero false positives (Fig. 3). When the DNA profile for only  
9 one parent exists in the database, the probability of a false nega-  
10 tive slightly increases and the power to unequivocally identify  
11 that parent slightly decreases (Vandeputte 2012; Vandeputte and  
12 Haffray 2014). The difference in statistical power between the  
13 duo and trio method stems from differences in statistical power  
14 that arise from the presence of SNP profiles for both parents  
15 (*TTR*) as opposed to one parent (*DTR*) in the reference database  
16 (Elston 1986; Goldgar and Thompson 1988). For a diploid (or  
17 allo-polyploid) organism genotyped with bi-allelic subgenome-  
18 specific DNA markers, two out of nine possible genotypic com-  
19 binations are informative for duo exclusion analysis, whereas  
20 12 out of 27 possible genotypic combinations are informative  
21 for trio exclusion analysis (Vandeputte 2012; Vandeputte and  
22 Haffray 2014). Moreover, trio exclusion analysis includes two  
23 highly informative (statistically powerful) combinations where  
24 the candidate offspring are heterozygous (*AB*) and both parents  
25 are homozygous for the same allele (either *AA* or *BB*).

26 Our computer forensic search did not recover pedigree  
27 records for 220 individuals in the UCD population; however, we  
28 suspected that their parents might be present in the SNP profile  
29 database. Using duo and trio exclusion analyses, we identified  
30 both parents for 214 individuals and one parent each for the  
31 other six individuals. Hence, using a combination of computer  
32 and DNA forensic approaches, 2,222 out of 2,470 possible par-  
33 ents of 1,235 individuals (90.0%) in the UCD population were  
34 identified and documented in the pedigree database (File S1;  
35 Fig. 2). The parents declared in pedigree records (if known),  
36 identified by DNA forensic methods (if conclusive), or both are  
37 documented in the pedigree database (File S1). Despite their  
38 historic and economic importance, the pedigrees of individuals  
39 preserved in the UCD Strawberry Germplasm Collection had not  
40 been previously documented. Besides reconstructing the geneal-  
41 ogy of the UCD population, previously hidden or unknown pedi-  
42 grees of extinct and extant individuals were discovered in the  
43 historic UCD records of Harold E. Thomas, Royce S. Bringham,  
44 and others (Bringham 1918-2016; Bringham *et al.* 1990; Johnson  
45 1990) and integrated into the global pedigree database (File S1).

46 To further validate the accuracy of DNA forensic approaches  
47 for parent identification in octoploid strawberry, we applied  
48 exclusion analysis to a population of 560 hybrid individuals  
49 developed from crosses among 30 UCD individuals (parents).  
50 The parents and hybrids and 1,561 additional UCD individuals  
51 were genotyped with 50K or 850K SNP arrays (Hardigan *et al.*  
52 2020a). The 50K array was developed with SNP markers from  
53 the 850K array (Hardigan *et al.* 2020a), which included a subset  
54 of 16,554 legacy SNP markers from the iStraw35 and iStraw90  
55 arrays (Bassil *et al.* 2015; Verma *et al.* 2016). We developed an  
56 integrated SNP profile database using 2,615 SNP markers common  
57 to the three arrays. Using parent-offspring trios, we discovered  
58 that the SNP profile for one of the parents (11C151P008) was  
59 a mismatch, whereas the SNP profiles of the other 29 parents  
60 perfectly matched their pedigree (birth) records. We discovered  
61 that the parent stated on the birth certificate for 11C151P008 was  
62 correct, but that the DNA sample and associated SNP marker

63 profile were incorrect. Hence, the DNA sample mismatch was  
64 traced by exclusion analysis to a single easily corrected labora-  
65 tory error.

66 These results highlight the power and accuracy of diploid  
67 Mendelian exclusion analysis for pedigree authentication (pa-  
68 ternity and maternity analysis), intellectual property protection,  
69 and quality control monitoring of germplasm and nursery stock  
70 collections in octoploid strawberry using subgenome-specific  
71 DNA markers. The application of these approaches was straight-  
72 forward because of the simplicity and accuracy of paralog- or  
73 homeolog-specific genotyping approaches in octoploid straw-  
74 berry populations (Hardigan *et al.* 2020a). The development and  
75 robustness of subgenome-specific genotyping approaches has  
76 enabled the application of standard diploid genetic theory and  
77 methods in octoploid strawberry, including the exclusion analy-  
78 sis methods applied in the present study (Jones and Ardren 2003;  
79 Vandeputte 2012; Vandeputte and Haffray 2014; Fig. 3). The  
80 power and accuracy of these methods were rigorously tested and  
81 affirmed in a court of law where DNA forensic evidence was piv-  
82 otal in proving the theft of University of California intellectual  
83 property (strawberry germplasm) by the defendants in a 2017  
84 case in US District Court for the Northern District of California  
85 captioned *The Regents of the University of California v California*  
86 *Berry Cultivars, LLC, Shaw, and Larson* (Chivvis 2017). The DNA  
87 forensic approach and evidence in that case are documented  
88 in a publicly available expert report identified by case number  
89 3:16-cv-02477 (<https://ecf.cand.uscourts.gov/cgi-bin/login.pl>).

### 90 **The Wild Roots of Cultivated Strawberry**

91 Our genealogy search did not uncover pedigree records for *F.*  
92  $\times$  *ananassa* cultivars developed between 1714 and 1775, the 61  
93 year period following the initial migration of *F. chiloensis* eco-  
94 types from Chile to Europe (Duchesne 1766; Darrow 1966). The  
95 scarcity of pedigree records from the eighteenth century was an-  
96 ticipated because the interspecific hybrid origin of *F. \times ananassa*  
97 was not discovered until mid-1700s (Duchesne 1766). ‘Madame  
98 Moutot’ was the only cultivar in the database with ancestry  
99 that could be directly traced to one of the putative original wild  
100 octoploid progenitors of the earliest *F. \times ananassa* hybrids that  
101 emerged in France in the early 1700s (Fig. 4). Although the  
102 genealogy primarily covers the last 200 years of domestication  
103 and breeding (File S1), ascendants in the pedigree of the culti-  
104 var ‘Madame Moutot’ (circa 1906) traced to ‘Chili de Plougastel’  
105 (Fig. 4), a putative clone of one of the original *F. chiloensis* subsp.  
106 *chiloensis* plants imported from Chile to France by the explorer  
107 Amédée-François Frézier (Gloede 1865; Carrière 1879; Bunyard  
108 1917; Darrow 1966; Pitrat and Fauray 2003). These plants were car-  
109 ried aboard the French frigate ‘St. Joseph’, delivered by Frézier  
110 to Brest, France (Bunyard 1917), and shared with Antoine Lau-  
111 rent de Jussieu, a botanist at the Jardin des plantes de Paris.  
112 According to de Lambertye (1864), the Frézier clone was widely  
113 disseminated and cultivated in Plougastel near Brest and inter-  
114 planted with *F. virginiana* (Duchesne 1766; Bunyard 1917; Pitrat  
115 and Fauray 2003). Hence, some of the earliest spontaneous hy-  
116 brids between *F. chiloensis* and *F. virginiana* undoubtedly arose  
117 in the strawberry fields of Brittany in the early 1700s (de Lam-  
118 bertye 1864; Darrow 1966; Pitrat and Fauray 2003). The French  
119 naturalist Bernard de Jussieu, the brother of Antoine Laurent de  
120 Jussieu and a mentor of Antoine Duchesne—“the father of the  
121 modern strawberry”—brought clones of the original Frézier *F.*  
122 *chiloensis* plants to the Jardins du Château de Versailles (Gardens  
123 of Versailles) where Duchesne (1766) unraveled the interspecific

1 hybrid origin of *F. × ananassa* (Darrow 1966; Williams 2001). The  
 2 next earliest *F. chiloensis* founders appear to be a California eco-  
 3 type identified in German breeding records from the mid-1800s  
 4 and an anonymous ecotype in the pedigree of the French cultivar  
 5 'La Constante' from 1855 (Files S1-S2; Gloede 1865; Merrick 1870;  
 6 Darrow 1937, 1966; Wilhelm and Sagen 1974).

7 The origins and identities of the earliest *F. virginiana* founders  
 8 of *F. × ananassa* remain a mystery because their migrations from  
 9 North America to Europe in the early 1600s and subsequent  
 10 intra-continental migrations were not well documented (File S1;  
 11 Duchesne 1766; de Lambertye 1864; Darrow 1937). The oldest  
 12 *F. virginiana* individuals identified in historic documents and  
 13 pedigree records were 'Large Early Scarlet' (1624), 'Old Scarlet'  
 14 (1625), and 'Hudson Bay' (1780), all extinct (File S1). We identi-  
 15 fied 30 anonymous *F. virginiana* and 76 anonymous *F. chiloensis*  
 16 founders in the pedigree records. These individuals were as-  
 17 signed unique alphanumeric aliases to facilitate reconstruction  
 18 of the genealogy, e.g., FV22 is the alias for an anonymous *F.*  
 19 *virginiana* founder and FC71 is the alias for an anonymous *F.*  
 20 *chiloensis* founder in the pedigree of 'Madame Moutot' (Fig. 4;

File S1).

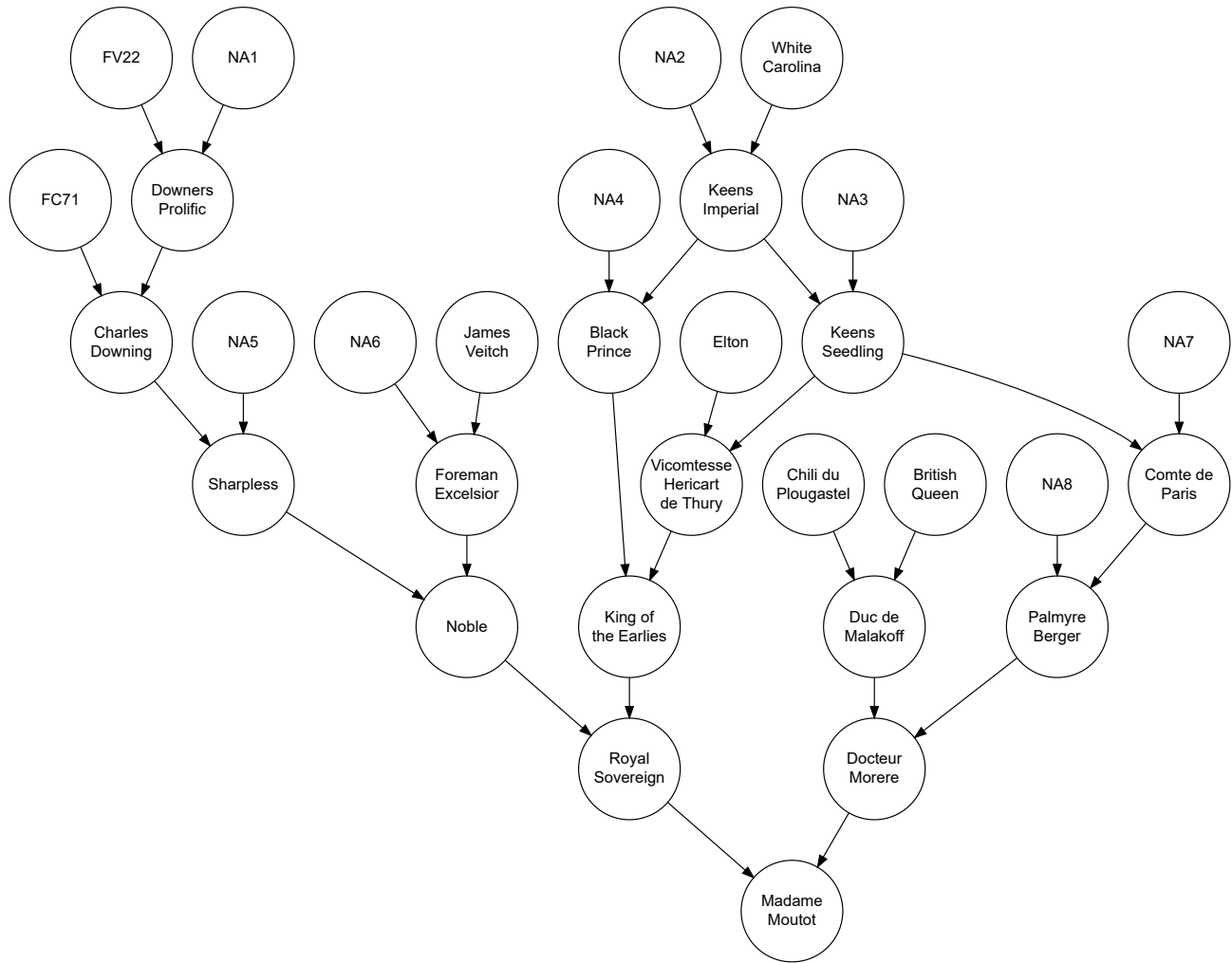
21

### The Complex Hybrid Ancestry of Cultivated Strawberry

22

23 Once the interspecific hybrid origin of *F. × ananassa* became  
 24 widely known (Duchesne 1766), domestication began in earnest  
 25 with extensive intra- and interspecific hybridization, artificial se-  
 26 lection, and intra- and intercontinental migration (Merrick 1870;  
 27 Fletcher 1917; Darrow 1937). These forces shaped the genetic  
 28 structure of the *F. × ananassa* populations that emerged in Europe  
 29 and North America and ultimately migrated around the globe  
 30 (Fletcher 1917; Darrow 1966; Sjulín and Dale 1987; Johnson 1990;  
 31 Sjulín 2006; Horvath *et al.* 2011; Sánchez-Sevilla *et al.* 2015; Hardi-  
 32 gan *et al.* 2018, 2020b). Over the next 250 years, horticulturalists  
 33 and plant breeders repeatedly tapped into the wild reservoir of  
 34 genetic diversity, especially wild octoploid taxa native to North  
 35 America (Fig. 1; Table 1). There are numerous narrative accounts  
 36 of what transpired, especially in Europe, North America, and  
 37 California (Clausen 1915; Darrow 1937, 1966; Sjulín and Dale  
 38 1987; Bringham *et al.* 1990; Dale and Sjulín 1990; Johnson 1990;  
 39 Hancock *et al.* 2001; Sjulín 2006; Hancock *et al.* 2010; Horvath

39



**Figure 4** Pedigree for the Heirloom Cultivar 'Madame Moutot' (circa 1906). Arrows indicate the flow of genes from parents to offspring. FV22 is an unknown *F. virginiana* ecotype, FC71 is an unknown *F. chiloensis* ecotype, and 'Chili du Plougastel' is purportedly one of the original *F. chiloensis* individuals imported by Amédée-François Frézier from Chile to France in 1714. Unknown parents of individuals in the pedigree are identified by NA1, NA2, ..., NA7. Terminal individuals in the pedigree are founders (individuals with unknown parents). The oldest *F. × ananassa* cultivar in the pedigree is 'White Carolina' (PI551681), which originated sometime before 1775.

1 *et al.* 2011; Sánchez-Sevilla *et al.* 2015; Hancock *et al.* 2018) but  
 2 none have painted a holistic picture of the complicated wild  
 3 ancestry and dynamic forces that shaped genetic diversity in *F.*  
 4  $\times$  *ananassa*.

5 We identified 1,438 founders in the genealogy of cultivated  
 6 strawberry (Fig. 1; Table 1; Files S1, S4-S5). Here and elsewhere,  
 7 ‘founders’ are individuals with unknown parents, whereas ‘an-  
 8 cestors’ are ascendants that may or may not be founders (Lacy  
 9 1989, 1995). The terminal nodes in the pedigree networks are  
 10 either founders or the youngest descendants in a pedigree (Figs.  
 11 1-2). Of the 1,438 founders, 267 were wild species and 1,171  
 12 were *F. \times ananassa* individuals (Fig. 1; Table 1). Because the *F. \times*  
 13 *ananassa* founders are either interspecific hybrids or descendants  
 14 of interspecific hybrids, the number of wild species founders  
 15 could exceed 268. One of the challenges we had with estimat-  
 16 ing the number of wild species founders was the anonymity  
 17 of ecotypes that were used as parents before breeders began  
 18 carefully documenting pedigrees (File S1). We could not rule  
 19 out that some of the anonymous wild species founders in the  
 20 pedigree records might have been clones of the same individ-

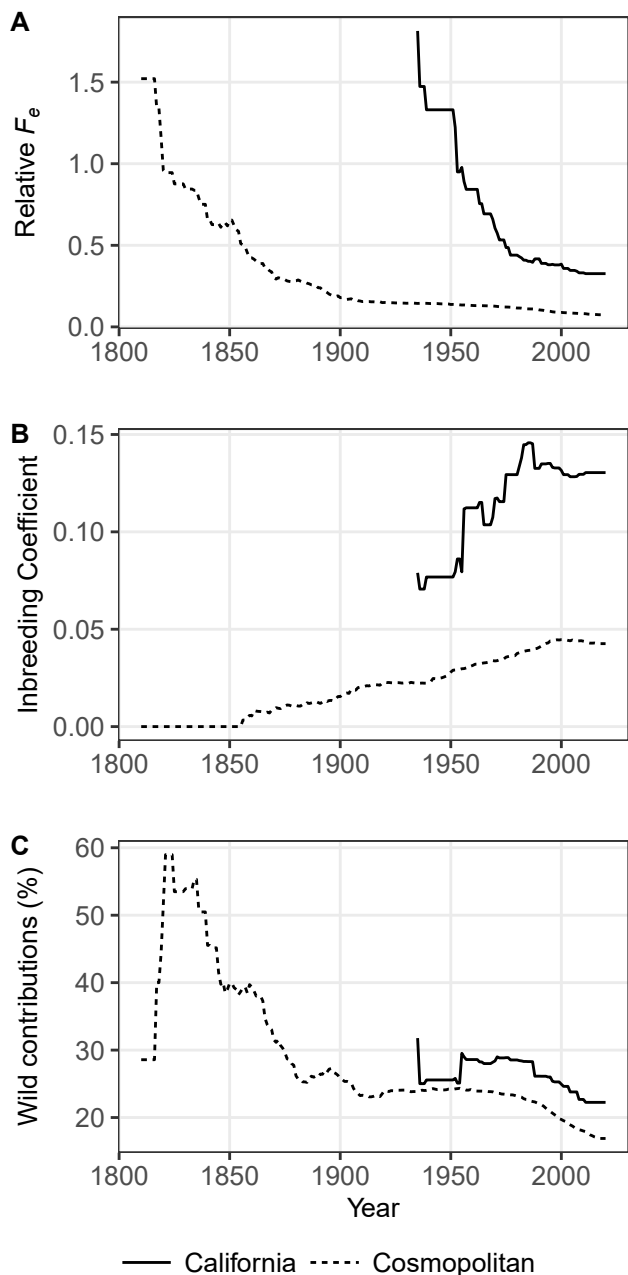
21 uals, which means that the estimated number of wild species  
 22 founders reported here could be inflated.

23 As interspecific hybridization with wild founders became  
 24 less important and intraspecific (*F. \times ananassa*) hybridization be-  
 25 came more important in strawberry breeding, the proportional  
 26 genetic contribution of wild founders to the gene pool of culti-  
 27 vated strawberry decreased (Fig. 5; Files S4-S5). This seems para-  
 28 doxical because 100% of the alleles found in *F. \times ananassa* were  
 29 inherited from wild founders, but increasingly flowed through *F.*  
 30  $\times$  *ananassa* descendants over time—wild octoploids numerically  
 31 only constituted 14% of the founders we identified (Table 1).  
 32 Several trends emerged from our analyses of genetic relation-  
 33 ships and founder contributions. First, inbreeding has steadily  
 34 increased over time as a consequence of population bottlenecks  
 35 and directional selection (Fig. 5B). Second, the California pop-  
 36 ulation was significantly more inbred than the cosmopolitan  
 37 population (Fig. 5B). These results were consistent with the find-  
 38 ings of Hardigan *et al.* (2020b) from genome-wide analyses of  
 39 DNA variants and population structure. They found selective  
 40 sweeps on several chromosomes in the California population,

**Table 1** Number of Primary, Secondary, and Tertiary Gene Pool Founders in the Global Genealogy of Cultivated Strawberry

Species	Ploidy	Giant	Halo	Complete
<b>Primary Gene Pool</b>				
<i>F. chiloensis</i>	2n = 8x = 56	79	33	112
<i>F. virginiana</i>	2n = 8x = 56	41	24	65
<i>F. \times ananassa</i>	2n = 8x = 56	656	515	1,171
Unknown Octoploid <i>Fragaria</i>	2n = 8x = 56	9	1	10
<b>Primary Gene Pool Total</b>		<b>785</b>	<b>573</b>	<b>1,358</b>
<b>Secondary Gene Pool</b>				
<i>F. iinumae</i>	2n = 2x = 14	1	2	3
<i>F. nilgerrensis</i>	2n = 2x = 14	2	0	2
<i>F. nipponica</i>	2n = 2x = 14	0	2	2
<i>F. nubicola</i>	2n = 2x = 14	2	0	2
<i>F. orientalis</i>	2n = 2x = 14	3	1	4
<i>F. viridis</i>	2n = 2x = 14	4	2	6
<i>F. vesca</i>	2n = 2x = 14	20	24	44
<i>F. moschata</i>	2n = 6x = 42	6	0	6
<i>F. \times vescana</i>	2n = 10x = 70	1	0	1
<b>Secondary Gene Pool Total</b>		<b>39</b>	<b>31</b>	<b>70</b>
<b>Tertiary Gene Pool</b>				
<i>P. glandulosa</i>	2n = 2x = 14	3	0	3
<i>P. anserina</i>	2n = 4x = 28	1	0	1
<i>P. palustris</i>	2n = 6x = 42	1	4	5
Unknown <i>Potentilla</i>	NA	0	1	1
<b>Tertiary Gene Pool Total</b>		<b>5</b>	<b>5</b>	<b>10</b>

Founders are individuals with unknown parents. The sociogram for the global genealogy consisted of ‘giant’ and ‘halo’ components. The giant component consisted of the highly interconnected mass of individuals in the sociogram (pedigree network), whereas the halo component consisted of orphans and other isolated individuals in small dead-end pedigrees that were disconnected from the giant component.



**Figure 5** Relative Founder Equivalents, Inbreeding Coefficients, and Wild Founder Genetic Contributions Over Time. (A) Relative founder equivalent ( $F_e/n$ ) estimates for California and cosmopolitan cultivars over time, where  $F_e$  = founder equivalents and  $n$  = number of founders. The California population included 69 cultivars developed at the University of California, Davis (UCD) since the inception of the breeding program in 1924. The birth year (year of origin) was known for all of the UCD cultivars. The cosmopolitan population included 2,140 cultivars with known birth years. (B) Wright's coefficient of inbreeding ( $F$ ) for individuals in the California and cosmopolitan populations over time.  $F$  was estimated from the relationship matrix (A). (C) Estimates of the genetic contributions of wild species founders to allelic diversity in the California and cosmopolitan populations.

1 which was shown to be unique and bottlenecked. Finally, the  
 2 relative number of founder equivalents (Lacy 1989, 1995) has

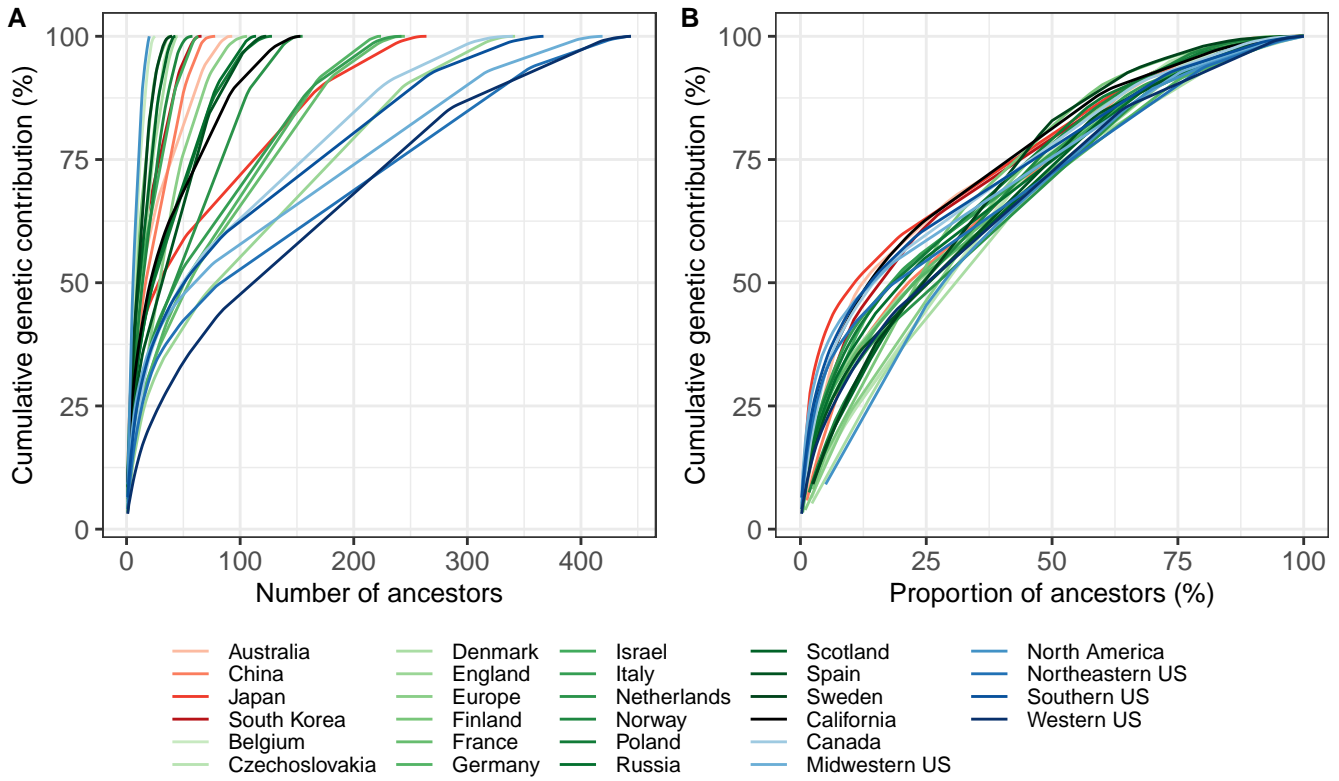
decreased over time, consistent with the increase in inbreeding  
 over time (Fig. 5A-B).

### Primary, Secondary, and Tertiary Gene Pool Founders of Cultivated Strawberry

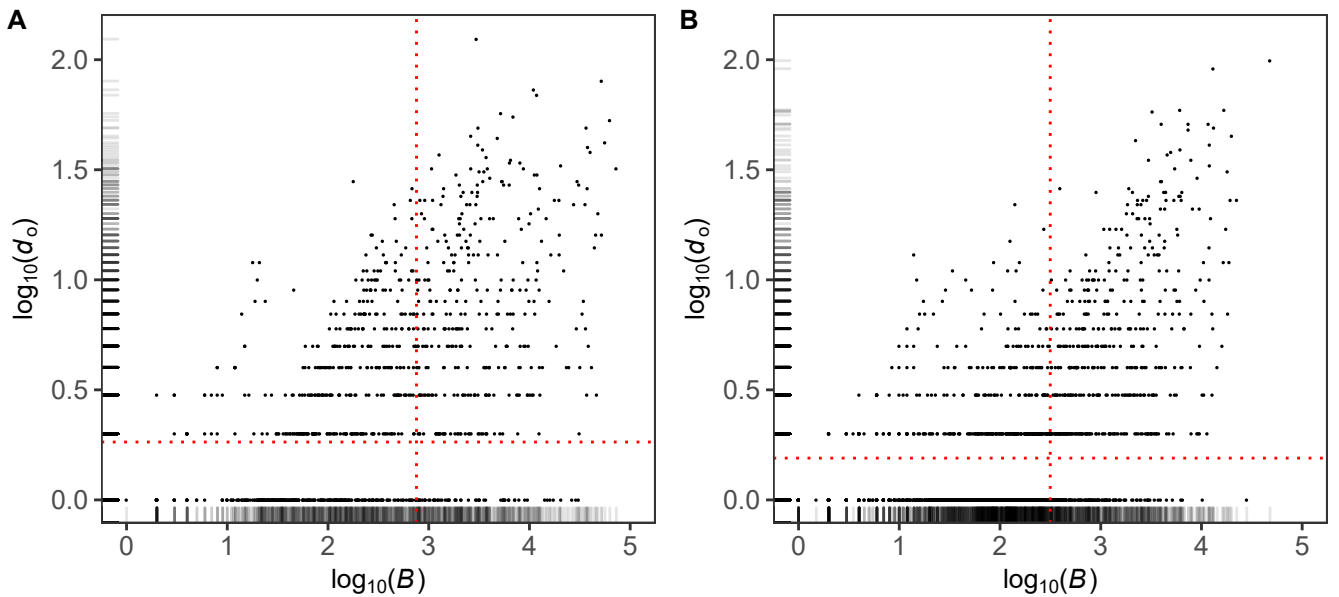
Predictably, the wild species founders of  $F. \times ananassa$  were dominated by  $F. chiloensis$  ( $n = 112$ ) and  $F. virginiana$  ( $n = 65$ ) (Table 1). Seven of eight subspecies of  $F. chiloensis$  and  $F. virginiana$  (Staudt 1988; Hummer et al. 2011) were identified in pedigree records:  $F. chiloensis$  subsp. *chiloensis*,  $F. chiloensis$  subsp. *lucida*,  $F. chiloensis$  subsp. *pacifica*, and  $F. chiloensis$  subsp. *sandwicensis*,  $F. virginiana$  subsp. *virginiana*,  $F. virginiana$  subsp. *glauca*, and  $F. virginiana$  subsp. *platyptala* (Bringhurst 1918-2016; <https://www.ars.usda.gov/>; Fig. 1; Table 1; File S1). Primary gene pool individuals (187 wild octoploid ecotypes and 1,171 hybrid  $F. \times ananassa$  individuals) constituted 95% of the founders and were estimated to have contributed  $\geq 99\%$  of the allelic diversity to global, California, and cosmopolitan  $F. \times ananassa$  populations (Fig. 6; Table 1; Files S4-S5). Even though wild species from the secondary ( $n = 70$ ) and tertiary ( $n = 10$ ) gene pools of  $F. \times ananassa$  constituted 6% of the founders and 30% of the wild species founders identified in pedigree records, they were estimated to have contributed  $< 0.1\%$  of the allelic diversity in the global  $F. \times ananassa$  population (Table 1; Files S4-S5).

The secondary and tertiary gene pool founders were primarily parents of orphans or other isolated individuals in short dead-end pedigrees that have not materially contributed allelic diversity to the primary gene pool. These included decaploid ( $2n = 10x = 70$ )  $F. \times vescana$  and pentaploid ( $2n = 5x = 35$ )  $F. \times bringhurstii$  individuals (Bringhurst and Senanayake 1966; Bauer 1994; Sangiacomo and Sullivan 1994; Hummer et al. 2011). Although frequently cited as important genetic resources for strawberry breeding (Darrow 1966; Hummer 2008; Gaston et al. 2020), the secondary and tertiary gene pools of cultivated strawberry have had limited utility because of the range of biological challenges one encounters when attempting to introgress alleles from exotic sources through interspecific and intergeneric hybrids, e.g., reproductive and recombination barriers, ploidy differences, meiotic abnormalities, and hybrid sterility (Bringhurst and Senanayake 1966; Bringhurst and Gill 1970; Harlan and de Wet 1971; Evans 1977; Bauer 1994; Sangiacomo and Sullivan 1994).

The secondary and tertiary gene pools are hardly needed to drive genetic gains or solve problems in strawberry breeding. Hardigan et al. (2020b) showed that genetic diversity is massive in the primary gene pool and has not been eroded by domestication and breeding on a global scale, even though it has been significantly reduced and restructured in certain populations, e.g., the California population. The profound changes and restructuring in the California population over time, as previously noted, were clearly evident in the sociograms and PCAs of the pedigree-genomic relationship matrices (Figs. 1-2). Because the California population has been the source of numerous historically and commercially important cultivars, we hypothesize that intense selection and population bottlenecks have purged a high frequency of unfavorable alleles compared to many other populations, thereby yielding an elite population with lower genetic diversity than the highly admixed cosmopolitan population (Figs. 1-2; Hardigan et al. 2020b).



**Figure 6** Genetic Contributions of Ancestors to Cultivars. (A) The genetic contributions of ancestors to the allelic diversity among  $k$  cultivars within a focal population were estimated from the mean coancestry between the  $i$ th ancestor and the  $k$  cultivars within the focal population. The genetic contributions of the ancestors were ordered from largest to smallest to calculate the cumulative genetic contributions of ancestors to cultivars in a focal population. (B) The proportion of ancestors needed to account for  $p\%$  of the allelic diversity among cultivars within a focal population was estimated by dividing the cumulative genetic contribution by  $k$ .



**Figure 7** Structural Roles and Betweenness Centrality ( $B$ ) and Out-Degree ( $d_o$ ) Statistics for Individuals in Cultivated Strawberry Sociograms. (A)  $B$  and  $d_o$  estimates for individuals in the California population. (B)  $B$  and  $d_o$  estimates for individuals in the cosmopolitan population. (A) and (B) The red dashed lines delineate globally central (upper right;  $d_o > \bar{d}_o \wedge B > \bar{B}$ ), locally central (upper left;  $d_o > \bar{d}_o \wedge B < \bar{B}$ ), broker (lower right;  $d_o < \bar{d}_o \wedge B > \bar{B}$ ), and marginal (lower left;  $d_o < \bar{d}_o \wedge B < \bar{B}$ ) quadrants, where  $\bar{B}$  = the mean of  $B$  estimates and  $\bar{d}_o$  = the mean of  $d_o$  estimates.  $\bar{B} = 755.6$  and  $\bar{d}_o = 1.8$  for the California population, whereas  $\bar{B} = 315.2$  and  $\bar{d}_o = 1.5$  for the cosmopolitan population.  $B$  and  $d_o$  estimate densities are plotted along the x- and y-axes.

**Table 2 The Twenty-Most Prominent and Historically Important Ancestors of Cultivars**

California				Cosmopolitan			
Ancestor	GC (%)	B	$d_o$	Ancestor	GC (%)	B	$d_o$
Tufts	12.2	52,013.9	80	Howard 17	4.4	47,942.5	99
Lassen	7.1	56,157.0	42	Fairfax	1.9	13,090.4	91
Cal 177.21	6.4	36,728.6	49	Hovey	1.8	12,390.6	19
Douglas	5.7	72,781.8	32	Tufts	1.4	16,579.3	12
71C098P605	3.6	16,434.8	13	Crescent	1.3	16,803.7	59
Nich Ohmer	3.0	2,977.0	124	Aberdeen	1.2	7,908.6	35
Camino Real	2.6	17,797.1	23	Sharpless	1.2	11,727.0	51
Howard 17	2.5	52,231.1	16	Blakemore	1.2	13,265.9	49
Sequoia	2.4	40,254.5	38	Wilson	1.0	4,012.6	51
Diamante	2.3	31,032.9	27	Royal Sovereign	0.9	19,373.0	23
Irvine	2.0	11,938.8	12	Harunoka	0.9	6,193.6	24
Palomar	1.9	27,644.3	22	Douglas	0.8	22,433.6	23
Albion	1.8	22,016.6	11	Gorella	0.7	12,053.2	41
42C008P016	1.8	12,687.4	26	Hoffman	0.7	5,738.0	17
Parker	1.5	2,924.8	10	Marshall	0.7	0.0	58
65C065P601	1.5	19,867.1	13	Holiday	0.6	6,157.4	39
Seascape	1.5	8,637.0	12	Senga Sengana	0.6	3,258.0	58
San Andreas	1.3	35,857.9	22	Bubach	0.6	0.0	56
Aiko	1.2	8,141.0	5	Reiko	0.6	2,766.0	19
Oso Grande	1.1	48,118.7	20	Cumberland Triumph	0.5	10,544.7	12

Genetic contribution statistics are tabulated for the twenty-most important ancestors of cultivars in the California and cosmopolitan populations. The proportional genetic contribution of the  $i$ th ancestor to cultivars within a population was estimated by  $P_i = GC_i / \sum_i GC_i$ , where  $GC_i$  is the genetic contribution of  $i$ th ancestor to cultivars in the focal population.  $B$  is the betweenness centrality estimate of the ancestor in the focal population.  $B = 0$  for founders and  $B > 0$  for non-founders. Out-degree ( $d_o$ ) is the number of descendants of the ancestor in the focal population.

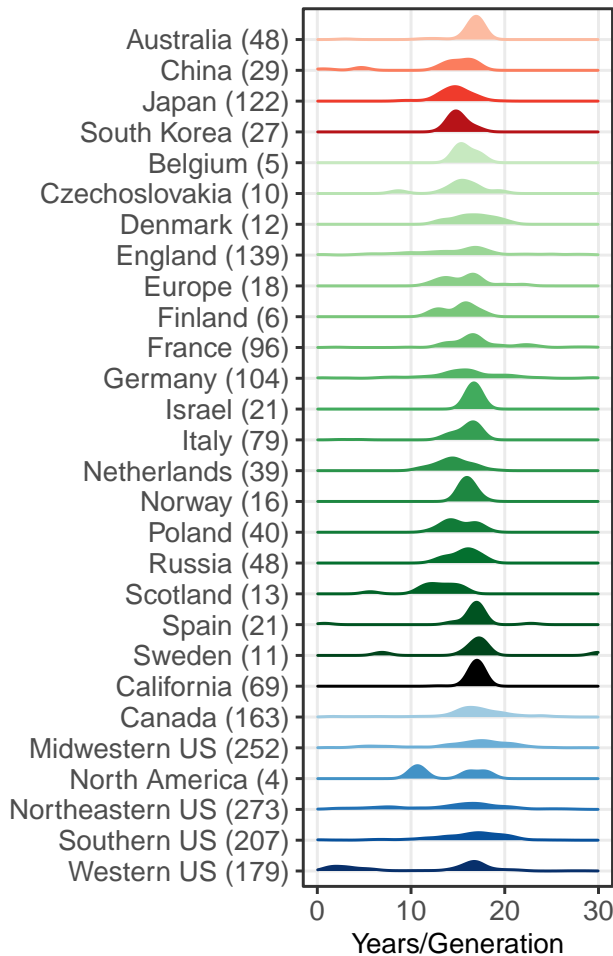
### 1 **Prominent and Historically Important Ancestors of Cultivated** 2 **Strawberry**

3 We used coancestry, betweenness centrality ( $B$ ), and out-degree  
4 ( $d_o$ ) statistics to estimate the genetic contribution (GC) of  
5 founders and non-founders to genetic variation within a popu-  
6 lation and identify the most prominent and important ancestors  
7 in the genealogy of cultivated strawberry (Freeman 1977; Scott  
8 1988; Lacy 1989, 1995; Fig. 6; Table 2; Files S4-S5). The estimation  
9 of GC from the coancestry matrix ( $A$ ) differed between founders  
10 and ancestors (founders and non-founders). For founders, GC  
11 was estimated by the mean coancestry or kinship ( $MK$ ) between  
12 each founder and cultivars within a focal population (Files S4-  
13 S5). For ancestors, GC was iteratively estimated by  $MK$  between  
14 each ancestor and cultivars within a focal population, starting  
15 with the ancestor with the largest  $MK$  estimated from  $A$ , deleting  
16 that ancestor, re-estimating the coancestry matrix ( $A^*$ ), selecting  
17 the ancestor with the largest  $MK$  estimated from the pruned  
18 coancestry matrix ( $A^*$ ), deleting that ancestor, re-estimating  
19 the coancestry matrix, and repeating until every ancestor had  
20 been dropped. We compiled GC,  $B$ , and  $d_o$  estimates for every  
21 founder and non-founder in the pedigree database (Files S4-S5).

We identified four *F. chiloensis*, five *F. virginiana*, and 40 *F. ×*  
*ananassa* founders in the genealogy of the California population  
22 (File S4). Cumulative GC estimates for the California population  
23 were 1.8% for *F. chiloensis*, 12.7% for *F. virginiana*, and 85.5% for  
24 *F. × ananassa* founders. Four of the nine wild octoploid founders  
25 of the California population were founders of the historic Etters-  
26 burg population that supplied genetic diversity for private and  
27 public sector breeding programs in California (Clausen 1915;  
28 Wilhelm and Sagen 1974; Bringhurst et al. 1990; Sjulín 2006). The  
29 wild octoploid founders with the largest genetic contributions  
30 were three *F. virginiana* ecotypes: 'New Jersey Scarlet' (8.3%),  
31 'Hudson Bay' (2.7%), and 'Wasatch' (1.3%) (Table S1). Wasatch is  
32 the *F. virginiana* subsp. *glauca* donor of the PERPETUAL FLOW-  
33 ERING mutation that Bringhurst et al. (1980) transferred into *F. ×*  
34 *ananassa* (Bringhurst et al. 1989). The Wasatch ecotype appears in  
35 the genetic background of every day-neutral cultivar developed  
36 at the University of California, Davis. Similarly, we identified 26  
37 *F. chiloensis*, 24 *F. virginiana*, and 490 *F. × ananassa* founders in the  
38 genealogy of the cosmopolitan population (File S5). Cumulative  
39 GC estimates for the cosmopolitan population were 4.6% for *F.*  
40 *chiloensis*, 14.1% for *F. virginiana*, 79.9% for *F. × ananassa*, and 1.4%  
41  
42

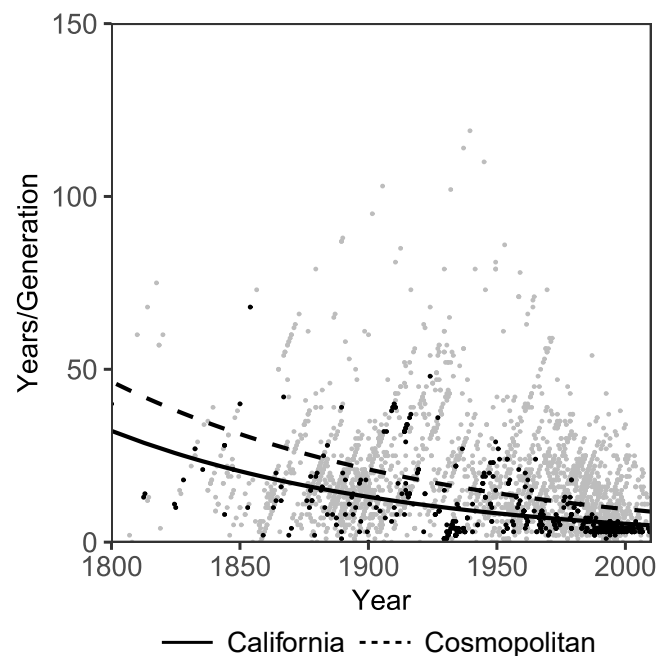
1 for other founders. Similar to what we found for the California  
 2 population, the wild octoploid founders with the largest genetic  
 3 contributions were 'New Jersey Scarlet' (8.3%) and 'Hudson Bay'  
 4 (3.5%) (Fletcher 1917; Darrow 1937). The next largest genetic  
 5 contribution was made by FC\_071 (1.9%), an *F. chiloensis* ecotype  
 6 of unknown origin found in the pedigrees of Madame Moutot,  
 7 Sharpless, Royal Sovereign, and other influential early cultivars  
 8 (Table S1; Figure 4).

9 A significant fraction of the alleles found in *F. × ananassa*  
 10 populations have flowed through a comparatively small number  
 11 of common ancestors, each of which have contributed unequally  
 12 to standing genetic variation (Fig. 6; Table 2; Files S4-S5). The



**Figure 8** Selection Cycle Length Distributions by Geography. Selection cycle length means ( $\bar{S}$  = mean number of years/generation) were estimated for  $k$  cultivars within continent-, region-, and country-specific focal populations of cultivated strawberry ( $k$  is shown in parentheses for each geographic group).  $\bar{S}$  was estimated from edge lengths (years/edge) for all possible paths (directed graphs with alleles flowing from parents to offspring but not *vice versa*) in pedigrees connecting cultivars to founders, where the length of an edge = the birth year difference between parent and offspring.  $\bar{S}$  probability densities are shown for cultivars developed in different countries, regions, or continents. Only estimates in the zero to 30 year/generation range are shown because estimates exceeding 30 years/generation were extremely rare.

13 most important ancestors are described as 'stars' in the lexicon  
 14 of social network analysis, and are either locally or globally  
 15 central (Moreno 1953; Scott 1988; Wasserman and Faust 1994).  
 16 Globally central individuals reside in the upper right quadrant  
 17 of the  $B \times d_o$  distribution ( $d_o > \bar{d}_o \wedge B > \bar{B}$ ), where  $\bar{B}$  is the  
 18 mean of  $B$  and  $\bar{d}_o$  is the mean of  $d_o$ —8.7-8.9% of the ancestors  
 19 were classified as globally central (Fig. 7; Moreno 1953; Scott  
 20 1988; Wasserman and Faust 1994). Locally central individuals  
 21 reside in the upper left quadrant of the  $B \times d_o$  distribution ( $d_o$   
 22  $> \bar{d}_o \wedge B < \bar{B}$ )—11.8-12.1% of the ancestors were classified as  
 23 locally central (Fig. 7; Moreno 1953; Scott 1988; Wasserman and  
 24 Faust 1994). 'Tufts', 'Lassen', 'Nich Ohmer', 'Howard 17', and  
 25 'Fairfax' were among the biggest stars, along with several other  
 26 iconic, mostly heirloom cultivars, and all were either locally  
 27 or globally central (Table 2). Stars are 'gatekeepers' that have  
 28 numerous descendants (the largest  $d_o$  estimates), transmitted a  
 29 disproportionate fraction of the alleles found in a population  
 30 (have the largest GC estimates), have the largest number of  
 31 inter-connections (largest  $B$  estimates) in the pedigree, and are  
 32 visible in sociograms as nodes with radiating pinwheel-shaped  
 33 patterns of lines (Fig 2; Table 2; Files S4-S5). Several of the latter  
 34 are visible in the sociograms we developed for the California  
 35 and cosmopolitan populations. Stars have the largest nodes ( $B$   
 36 estimates) in the sociograms (Fig 2).



**Figure 9** Breeding Speed Over Time. Selection cycle lengths ( $S$  = years/generation) were estimated for 3,693 independent parent-offspring edges in the pedigree networks for the California and cosmopolitan populations.  $S$  estimates were limited to parents and offspring with known birth years. Selection cycle lengths are plotted against the midpoint ( $m$ ) between parent and offspring birth years for California (black points) and cosmopolitan (gray points) populations. The plotted lines are exponential decay functions fitted by non-linear regression of  $S$  on  $m$ . The function for the California population was  $y = 35.06 \cdot e^{-0.0090 \cdot (x-1790.5)}$  (Nagelkerke pseudo- $R^2 = 0.25$ ;  $p < 0.001$ ). The function for the cosmopolitan population was  $y = 76.69 \cdot e^{-0.0079 \cdot (x-1736.5)}$  (Nagelkerke pseudo- $R^2 = 0.08$ ;  $p < 0.001$ ).

We estimated and compiled GC statistics for every ancestor in the California and cosmopolitan populations (Files S4-S5). The twenty-most prominent and historically important ancestors of the California and cosmopolitan populations are shown in Table 2. They include several iconic and well known heirloom and modern cultivars, e.g., 'Tioga', 'Douglas', and 'Royal Sovereign' (Fletcher 1917; Darrow 1937, 1966; Wilhelm and Sagen 1974; Sjulín and Dale 1987; Bringhurst *et al.* 1990), in addition to 'unreleased' germplasm accessions preserved in the UCD Strawberry Germplasm Collection, e.g., 65C065P601 (aka 65.65-1). The latter is the oldest living descendant of the aforementioned *F. virginiana* subsp. *glauca* 'Wasatch' ecotype collected by Royce S. Bringhurst from the Wasatch Mountains, Little Cottonwood Canyon, Utah (Bringhurst *et al.* 1980, 1989; Ahmadi *et al.* 1990). The 'Wasatch' ecotype is a founder of every day-neutral cultivar in the California population and many day-neutral cultivars in the cosmopolitan population with alleles flowing through 65C065P601 and the UCD cultivar 'Selva' (Bringhurst *et al.* 1989; Files S4-S5).

GC statistics were ordered from largest to smallest ( $GC_1 \geq GC_2 \geq \dots \geq GC_n$ ) and progressively summed to calculate the cumulative genetic contributions of ancestors and the number of ancestors needed to explain  $p\%$  of the genetic variation ( $n_p$ ) in a focal population, where  $p$  ranges from 0 to 100% (Fig. 6). The parameter  $n_{100}$  estimates the number of ancestors needed to account for 100% of the genetic variation among  $k$  cultivars in a focal population (each focal population was comprised of cultivars, ascendants, and descendants).  $n_{100}$  estimates were 153 for the California population and 3,240 for the cosmopolitan population. The latter number was significantly larger than the number for the California population because the cosmopolitan population includes pedigrees for 2,499 cultivars developed worldwide, whereas the California population includes pedigrees for 69 UCD cultivars only (File S1). Within European countries,  $n_{100}$  ranged from 25 for Belgium to 342 for England (Fig. 6A). Within the US,  $n_{100}$  ranged from a minimum of 367 for the southern region to a maximum of 444 for western and northeastern regions.

Predictably,  $n_p$  increased at a decreasing rate as the number of GC-ranked ancestors increased (Fig. 6). Cumulative GC estimates increased as non-linear diminishing-return functions of the number of ancestors (Table 2; Files S4-S5). The slopes were initially steep because a fairly small number of ancestors accounted for a large fraction of the genetic variation within a particular focal population. Across continents, regions, and countries, eight to 112 ancestors accounted for 50% of the allelic variation within focal populations (Fig. 6; Table 2). The differences in  $n_p$  estimates were partly a function of the number of cultivars ( $k$ ) within each focal population. When  $n_p$  was expressed as a function of  $k$ , we found that the proportion of ancestors needed to explained  $p\%$  of the allelic variation in a focal population was strikingly similar across continents, regions, and countries, e.g., the Western US population, which had the largest  $n_{100}$  estimate (Fig. 6A), fell squarely in the middle when expressed as a function of  $k$  (Fig. 6B).

### Breeding Speed in Cultivated Strawberry

Social network analyses of the pedigree networks shed light on the speed of breeding and changes in the speed of breeding over the last 200 years in strawberry (Figs. 8-9). We retraced the ancestry of every cultivar through nodes and edges in the sociograms (Figs. 1-2). The year of origin was known for 71%

of the individuals. These edges yielded robust estimates of the mean selection cycle length in years ( $\bar{S}$  = mean number of years/generation).  $\bar{S}$  was calculated from thousands of directed acyclic graphs, which are unidirectional paths traced from cultivars back through descendants to founders (Thulasiraman and Swamy 1992). Collectively, cultivars in the California population ( $n = 69$ ) visited 27,058 parent-offspring edges, whereas cultivars in the cosmopolitan population ( $n = 1,982$ ) visited 155,487 parent-offspring edges. The selection cycle length means ( $\bar{S}$ ) and distributions over the last 200 years were strikingly similar across continents, regions, and countries— $\bar{S}$  was 16.9 years/generation for the California population and 16.0 years/generation for the cosmopolitan population (Fig. 8). These extraordinarily long selection cycle lengths are more typical of a long-lived woody perennial than a fast cycling annual (van Nocker and Gardiner 2014; Jighly *et al.* 2019); however, the speed of breeding has steadily increased over time (Fig. 8). By 2000,  $\bar{S}$  had decreased to six years/generation in the California population and 10 years/generation in the cosmopolitan population (Fig. 9).

The genealogy does not account for lineages underlying what must have been millions of hybrid progeny screened in breeding programs worldwide, e.g., Johnson (1990) alone reported screening 600,000 progeny over 34 years (1956-1990) at Driscoll's (Watsonville, California). Cultivars are nevertheless an accurate barometer of global breeding activity and the only outward facing barometer of progress in strawberry breeding. When translated across the last 200 years of breeding, our selection cycle length estimates imply that the 2,656 cultivars in the genealogy of cultivated strawberry have emerged from the mathematical equivalent of only 12.9 cycles of selection (200 years  $\div$  15.5 years per generation). Even though offspring from 250 years of crosses have undoubtedly been screened worldwide since 1770, 15.5 years have elapsed on average between parents and offspring throughout the history of strawberry breeding (Fig. 8-9). Because genetic gains are affected by selection cycle lengths, and faster generation times normally translate into greater genetic gains and an increase in the number of recombination events per unit of time (Bernardo 2002; Ceccarelli 2015; Bernardo 2017; Jighly *et al.* 2019; Bernardo 2020), our analyses suggest that genetic gains can be further increased in strawberry by shortening selection cycle lengths. Genome-informed breeding, speed breeding, and other technical innovations are geared towards that goal and have the potential to shorten selection cycle lengths and increase genetic gains (van Nocker and Gardiner 2014; Whitaker *et al.* 2020).

### Acknowledgements

This research was supported by grants to SJK from the United States Department of Agriculture (<http://dx.doi.org/10.13039/100000199>) National Institute of Food and Agriculture (NIFA) Specialty Crops Research Initiative (#2017-51181-26833), California Strawberry Commission (<http://dx.doi.org/10.13039/100006760>), and the University of California, Davis (<http://dx.doi.org/10.13039/100007707>) and to SLD from the National Science Foundation Division Of Integrative Organismal Systems (#1444478). The USDA grant supported the dissertation research of DDP and MJF. The postdoctoral research of CH was supported by the NSF grant. We are grateful to Clint Pumphrey, the manuscript curator of the special collections and archives of the Merrill-Cazier Library at Utah State University (Logan, Utah). Clint assisted the first



1 author with acquiring and researching the laboratory notebooks  
 2 and other records of Royce S. Bringhurst (1918-2005), a former  
 3 faculty member and strawberry breeder at the University of  
 4 California, Davis (1953-1989). The documents and photos  
 5 associated with the collection yielded extensive pedigree  
 6 records that were crucial for reconstructing the genealogy  
 7 of the UCD Strawberry Breeding Program. We are equally  
 8 grateful to Phillip Stewart, a strawberry breeder at Driscoll's  
 9 (Watsonville, California), for sharing copies of the University  
 10 of California, Berkeley (UCB) pedigree records of Harold E.  
 11 Thomas (1900-1986), a former faculty member and strawberry  
 12 breeder at UCB from 1927 to 1945. Those pedigree records  
 13 greatly increased the completeness and depth of the database  
 14 for the early years of the University of California Strawberry  
 15 Breeding Program. The authors thank Thomas Sjulín, a former  
 16 strawberry breeder at Driscoll's (Watsonville, California), for  
 17 sharing the public pedigree records he assembled over his career.  
 18 Those nucleated the pedigree database we developed and were  
 19 a catalyst for our study. SJK and GSC thank Robert Kerner (In-  
 20 formation Technology Manager, Department of Plant Sciences,  
 21 UCD) for the computer forensic analyses that recovered several  
 22 hundred pedigree records for UCD individuals from an obsolete  
 23 electronic database, thus preventing the loss of those records  
 24 for perpetuity. They were critical for integrating the UCD  
 25 genealogy with the global genealogy for cultivated strawberry.  
 26 SJK especially thanks Rachel Krevans, Matthew Chivvus, Jake  
 27 Ewert, and Wesley Overson (lawyers at Morrison-Forester,  
 28 San Francisco, California) for their integrity, friendship, and  
 29 steadfast support.

### 30 Literature Cited

31 Affymetrix Inc., 2015 Axiom® genotyping solution data analysis  
 32 guide (P/N 702961 Rev. 3). Affymetrix, Inc., Santa Clara, CA.  
 33 Ahmadi, H., R. S. Bringhurst, and V. Voth, 1990 Modes of inher-  
 34 itance of photoperiodism in *fragaria*. *J. Am. Soc. Hortic. Sci.*  
 35 **115**: 146–152.  
 36 Barabási, A.-L., 2016 *Network science*. Cambridge University  
 37 Press, Cambridge, England.  
 38 Barabási, A.-L., N. Gulbahce, and J. Loscalzo, 2011 Network  
 39 medicine: a network-based approach to human disease. *Nat.*  
 40 *Rev. Genet.* **12**: 56–68.  
 41 Bassil, N. V., T. M. Davis, H. Zhang, S. Ficklin, M. Mittmann,  
 42 *et al.*, 2015 Development and preliminary evaluation of a 90K  
 43 Axiom® SNP array for the allo-octoploid cultivated straw-  
 44 berry *Fragaria* × *ananassa*. *BMC Genomics* **16**: 155.  
 45 Bastian, M., S. Heymann, and M. Jacomy, 2009 Gephi: an open  
 46 source software for exploring and manipulating networks. In  
 47 *Third international AAAI conference on weblogs and social media*,  
 48 pp. 361–362.  
 49 Bauer, A., 1994 Progress in breeding decaploid *Fragaria* × *vescana*  
 50 hybrids. In *Progress in Temperate Fruit Breeding*, pp. 189–191,  
 51 Springer, Dordrecht, Netherlands.  
 52 Bernardo, R., 2002 *Breeding for quantitative traits in plants*. Stemma  
 53 Press, Woodbury, MN.  
 54 Bernardo, R., 2017 Prospective targeted recombination and ge-  
 55 netic gains for quantitative traits in maize. *Plant Genome* **10**.  
 56 Bernardo, R., 2020 Reinventing quantitative genetics for plant  
 57 breeding: something old, something new, something bor-  
 58 rowed, something BLUE. *Heredity* pp. 1–11.  
 59 Berry, K. J., J. E. Johnston, and P. W. Mielke Jr, 2014 *A chronicle of*  
 60 *permutation statistical methods*. Springer, Cham, Switzerland.

Brandes, U., 2001 A faster algorithm for betweenness centrality. *J. Math. Sociol.* **25**: 163–177.  
 Bringhurst, R., H. Ahmadi, and V. Voth, 1989 Inheritance of the day-neutral trait in strawberries. *Acta Hort.* **265**: 35–42.  
 Bringhurst, R. and T. Gill, 1970 Origin of *Fragaria* polyploids. ii. unreduced and doubled-unreduced gametes. *Am. J. Bot.* **57**: 969–976.  
 Bringhurst, R. and Y. Senanayake, 1966 The evolutionary significance of natural *Fragaria chiloensis* × *F. vesca* hybrids resulting from unreduced gametes. *Am. J. Bot.* **53**: 1000–1006.  
 Bringhurst, R., V. Voth, *et al.*, 1980 Six new strawberry varieties released. *Calif. Agric.* **34**: 12–15.  
 Bringhurst, R. S., 1918-2016 Royce S. Bringhurst papers, 1918-2016. USU\_COLL MSS 515. Merrill-Cazier Library Special Collections & Archives, Utah State University, Logan, UT. <http://archiveswest.orbiscascade.org/ark:/80444/xv47241>.  
 Bringhurst, R. S., V. Voth, and D. Shaw, 1990 University of California strawberry breeding. *HortScience* **25**: 834–999.  
 Bunyard, E. A., 1917 The history and development of the strawberry. *J. Int. Garden Club* **1**: 69–90.  
 Carrière, E.-A., 1879 Fraisier du Chili. *Revue Horticole* **51**: 110–112.  
 Ceccarelli, S., 2015 Efficiency of plant breeding. *Crop Sci.* **55**: 87–97.  
 Chakraborty, R., M. Shaw, and W. J. Schull, 1974 Exclusion of paternity: the current state of the art. *Am. J. Hum. Genet.* **26**: 477.  
 Chivvis, M. A., 2017 The Regents of the University of California v California Berry Cultivars, LLC, Shaw, and Larson. *Intellectual Property Magazine* **November 2017**.  
 Christensen, O. F., 2012 Compatibility of pedigree-based and marker-based relationship matrices for single-step genetic evaluation. *Genet. Sel. Evol.* **44**: 37.  
 Christensen, O. F., P. Madsen, B. Nielsen, T. Ostensen, and G. Su, 2012 Single-step methods for genomic evaluation in pigs. *Animal* **6**: 1565–1571.  
 Clausen, R. E., 1915 Ettersburg strawberries: successful hybridizing of many species and varieties in northern California leads to production of new sorts which are apparently adapted to meeting almost all requirements. *J. Hered.* **6**: 324–331.  
 Contandriopoulos, D., C. Larouche, M. Breton, and A. Brousselle, 2018 A sociogram is worth a thousand words: proposing a method for the visual analysis of narrative data. *Qual. Res.* **18**: 70–87.  
 Cornille, A., T. Giraud, M. J. Smulders, I. Roldán-Ruiz, and P. Gladieux, 2014 The domestication and evolutionary ecology of apples. *Trends Genet.* **30**: 57–65.  
 Csardi, G. and T. Nepusz, 2006 The igraph software package for complex network research. *InterJournal Complex Systems*: 1695.  
 Dale, A. and T. M. Sjulín, 1990 Few cytoplasm contribute to North American strawberry cultivars. *HortScience* **25**: 1341–1342.  
 Darrow, G. M., 1937 Strawberry improvement. In *United States Department of Agriculture Yearbook of Agriculture*, pp. 445–495, United States Government Printing Office, Washington, D.C.  
 Darrow, G. M., 1966 *The Strawberry: history, breeding and physiology*. Holt, Rinehart & Winston, New York, NY.  
 de Lambertye, L., 1864 *Le Fraisier: sa botanique, son histoire, sa culture*. Librairie Centrale d'Agriculture et de Jardinage, Paris, France.  
 Diez, C. M., I. Trujillo, N. Martínez-Urdiroz, D. Barranco,

- 1 L. Rallo, *et al.*, 2015 Olive domestication and diversification in  
2 the Mediterranean Basin. *New Phytol.* **206**: 436–447.
- 3 Dillenberger, M. S., N. Wei, J. A. Tennesen, T.-L. Ashman, and  
4 A. Liston, 2018 Plastid genomes reveal recurrent formation of  
5 allopolyploid *Fragaria*. *Am. J. Bot.* **105**: 862–874.
- 6 Duan, N., Y. Bai, H. Sun, N. Wang, Y. Ma, *et al.*, 2017 Genome  
7 re-sequencing reveals the history of apple and supports a two-  
8 stage model for fruit enlargement. *Nat. Commun.* **8**: 1–11.
- 9 Duchesne, A.-N., 1766 *Histoire naturelle des fraisiers*. Didot le  
10 Jeune et C. J. Panckoucke, Paris, France.
- 11 Edwards, A., 1992 The structure of the Polar Eskimo genealogy.  
12 *Hum. Hered.* **42**: 242–252.
- 13 Efron, B., 1982 *The jackknife, the bootstrap, and other resampling*  
14 *plans*, volume 38. Siam, Philadelphia, PA.
- 15 Elston, R., 1986 Probability and paternity testing. *Am. J. Hum.*  
16 *Genet.* **39**: 112.
- 17 Endelman, J. B., 2011 Ridge regression and other kernels for  
18 genomic selection with R package rrBLUP. *Plant Genome* **4**:  
19 250–255.
- 20 Evans, W., 1977 The use of synthetic octoploids in strawberry  
21 breeding. *Euphytica* **26**: 497–503.
- 22 Finn, C. E., J. B. Retamales, G. A. Lobos, and J. F. Hancock, 2013  
23 The Chilean strawberry (*Fragaria chiloensis*): Over 1000 years  
24 of domestication. *HortScience* **48**: 418–421.
- 25 Fletcher, S. W., 1917 *The Strawberry in North America: history,*  
26 *origin, botany, and breeding*. The Macmillan Company, New  
27 York, NY.
- 28 Fradgley, N., K. A. Gardner, J. Cockram, J. Elderfield, J. M.  
29 Hickey, *et al.*, 2019 A large-scale pedigree resource of wheat  
30 reveals evidence for adaptation and selection by breeders. *PLoS*  
31 *Biol.* **17**: e3000071.
- 32 Freeman, L. C., 1977 A set of measures of centrality based on  
33 betweenness. *Sociometry* pp. 35–41.
- 34 Gao, H., O. F. Christensen, P. Madsen, U. S. Nielsen, Y. Zhang,  
35 *et al.*, 2012 Comparison on genomic predictions using three  
36 GBLUP methods and two single-step blending methods in the  
37 Nordic Holstein population. *Genet. Sel. Evol.* **44**: 8.
- 38 Gaston, A., S. Osorio, B. Denoyes, and C. Rothan, 2020 Applying  
39 the solanaceae strategies to strawberry crop improvement.  
40 *Trends Plant Sci.* **25**: 130–140.
- 41 Gloede, F., 1865 *Les bonnes fraises*. Librairie centrale d’agriculture  
42 et de jardinage, Paris, France.
- 43 Goldgar, D. and E. Thompson, 1988 Bayesian interval estimation  
44 of genetic relationships: application to paternity testing. *Am.*  
45 *J. Hum. Genet.* **42**: 135.
- 46 Graham, J., R. McNicol, and J. McNicol, 1996 A comparison of  
47 methods for the estimation of genetic diversity in strawberry  
48 cultivars. *Theor. Appl. Genet.* **93**: 402–406.
- 49 Gursoy, A., O. Keskin, and R. Nussinov, 2008 Topological prop-  
50 erties of protein interaction networks from a structural per-  
51 spective. *Biochem. Soc. Trans.* **36**: 1398–1403.
- 52 Hancock, J., P. Callow, A. Dale, J. Luby, C. Finn, *et al.*, 2001 From  
53 the Andes to the Rockies: Native strawberry collection and  
54 utilization. *HortScience* **36**: 221–225.
- 55 Hancock, J. and J. Luby, 1995 Adaptive zones and ancestry of  
56 the most important north american strawberry cultivars. *Fruit*  
57 *Var. J.* **49**: 85–90.
- 58 Hancock, J. F., P. P. Edger, P. W. Callow, T. Herlache, and C. E.  
59 Finn, 2018 Generating a unique germplasm base for the breed-  
60 ing of day-neutral strawberry cultivars. *HortScience* **53**: 1069–  
61 1071.
- 62 Hancock, J. F., C. E. Finn, J. J. Luby, A. Dale, P. W. Callow, *et al.*,  
2010 Reconstruction of the strawberry, *Fragaria* × *ananassa*,  
using genotypes of *F. virginiana* and *F. chiloensis*. *HortScience*  
**45**: 1006–1013.
- Hancock, J. F., T. Sjulín, and G. Lobos, 2008 Strawberries. In  
*Temperate Fruit Crop Breeding*, edited by J. F. Hancock, pp. 393–  
437, Springer, Dordrecht, Netherlands.
- Hardigan, M. A., M. J. Feldmann, A. Lorant, K. A. Bird, R. Fa-  
mula, *et al.*, 2020a Genome synteny has been conserved among  
the octoploid progenitors of cultivated strawberry over mil-  
lions of years of evolution. *Front. Plant Sci.* **10**: 1789.
- Hardigan, M. A., A. Lorant, D. D. A. Pincot, R. A. Famula, C. B.  
Acharya, *et al.*, 2020b Unraveling the complex hybrid ancestry  
and domestication history of cultivated strawberry. *Mol. Biol.*  
*Evol.* **Submitted, in review**.
- Hardigan, M. A., T. J. Poorten, C. B. Acharya, G. S. Cole, K. E.  
Hummer, *et al.*, 2018 Domestication of temperate and coastal  
hybrids with distinct ancestral gene selection in octoploid  
strawberry. *Plant Genome* **11**.
- Harlan, J. R. and J. M. de Wet, 1971 Toward a rational classifica-  
tion of cultivated plants. *Taxon* **20**: 509–517.
- Hartl, D. and A. Clark, 2007 *Principles of population genetics*. Sin-  
auer Associates, Sunderland, MA, fourth edition.
- Hayes, B., 2000 Computing science: Graph theory in practice:  
Part II. *Am. Sci.* **88**: 104–109.
- Horvath, A., J. Sánchez-Sevilla, F. Punelli, L. Richard, R. Sesmero-  
Carrasco, *et al.*, 2011 Structured diversity in octoploid straw-  
berry cultivars: importance of the old European germplasm.  
*Ann. Appl. Biol.* **159**: 358–371.
- Hummer, K., 2008 *Global conservation strategy for Fragaria (Straw-*  
*berry)*. International Society for Horticultural Science, Gent-  
Oostakker, Belgium.
- Hummer, K. E., N. Bassil, and W. Njuguna, 2011 *Fragaria*. In  
*Wild crop relatives: genomic and breeding resources*, pp. 17–44,  
Springer, Berlin, Germany.
- Jighly, A., Z. Lin, L. W. Pembleton, N. O. Cogan, G. C. Spangen-  
berg, *et al.*, 2019 Boosting genetic gain in allogamous crops  
via speed breeding and genomic selection. *Front. Plant Sci.* **10**:  
1364.
- Johnson, H. A., 1990 The contributions of private strawberry  
breeders. *HortScience* **25**: 897–902.
- Jones, A. G. and W. R. Ardren, 2003 Methods of parentage analy-  
sis in natural populations. *Mol. Ecol.* **12**: 2511–2523.
- Kim, H. and J. Song, 2013 Social network analysis of patent  
infringement lawsuits. *Technol. Forecast. Soc. Change* **80**: 944–  
955.
- Kominakis, A. P., 2001 Graph analysis of animals’ pedigrees.  
*Arch. Anim. Breed.* **44**: 521–530.
- Koschützki, D. and F. Schreiber, 2008 Centrality analysis meth-  
ods for biological networks and their application to gene regu-  
latory networks. *Gene Regul. Syst. Biol.* **2**: 193–201.
- Lacy, R. C., 1989 Analysis of founder representation in pedigrees:  
founder equivalents and founder genome equivalents. *Zoo*  
*Biol.* **8**: 111–123.
- Lacy, R. C., 1995 Clarification of genetic terms and their use in  
the management of captive populations. *Zoo Biol.* **14**: 565–577.
- Larson, G., D. R. Piperno, R. G. Allaby, M. D. Purugganan, L. An-  
dersson, *et al.*, 2014 Current perspectives and the future of  
domestication studies. *Proc. Natl. Acad. Sci. U.S.A* **111**: 6139–  
6146.
- Legarra, A., I. Aguilar, and I. Misztal, 2009 A relationship matrix  
including full pedigree and genomic information. *J. Dairy Sci.*  
**92**: 4656–4663.

- 1 Liston, A., R. Cronn, and T.-L. Ashman, 2014 *Fragaria*: a genus  
2 with deep historical roots and ripe for evolutionary and eco-  
3 logical insights. *Am. J. Bot.* **101**: 1686–1699.
- 4 Lynch, M. and B. Walsh, 1998 *Genetics and Analysis of Quantitative*  
5 *Traits*, volume 1. Sinauer, Sunderland, MA.
- 6 Mäkinen, V.-P., M. Parkkonen, M. Wessman, P.-H. Groop, T. Kan-  
7 ninen, *et al.*, 2005 High-throughput pedigree drawing. *Eur. J.*  
8 *Hum. Genet.* **13**: 987–989.
- 9 Manly, B. F., 2006 *Randomization, bootstrap and Monte Carlo meth-*  
10 *ods in biology*, volume 70. CRC press, Boca Raton, FL.
- 11 Merrick, J. M., 1870 *The Strawberry, and its Culture: with a descrip-*  
12 *tive catalogue of all known varieties*. J. E. Tilton and Company,  
13 Boston, MA.
- 14 Meyer, R. S., A. E. DuVal, and H. R. Jensen, 2012 Patterns and  
15 processes in crop domestication: an historical review and  
16 quantitative analysis of 203 global food crops. *New Phytol.*  
17 **196**: 29–48.
- 18 Meyer, R. S. and M. D. Purugganan, 2013 Evolution of crop  
19 species: genetics of domestication and diversification. *Nat.*  
20 *Rev. Genet.* **14**: 840–852.
- 21 Moreno, J. L., 1953 *Who shall survive? Foundations of sociometry,*  
22 *group psychotherapy and socio-drama*. Beacon House, Beacon,  
23 NY.
- 24 Morselli, C., 2010 Assessing vulnerable and strategic positions  
25 in a criminal network. *J. Contemp. Crim. Justice* **26**: 382–392.
- 26 Muranty, H., C. Denancé, L. Feugey, J.-L. Crépin, Y. Barbier, *et al.*,  
27 2020 Using whole-genome SNP data to reconstruct a large  
28 multi-generation pedigree in apple germplasm. *BMC Plant*  
29 *Biol.* **20**: 1–18.
- 30 Myles, S., A. R. Boyko, C. L. Owens, P. J. Brown, F. Grassi, *et al.*,  
31 2011 Genetic structure and domestication history of the grape.  
32 *Proc. Natl. Acad. Sci. U.S.A.* **108**: 3530–3535.
- 33 Nerghe, A., J.-S. Lee, P. Groenewegen, and I. Hellsten, 2015 Map-  
34 ping discursive dynamics of the financial crisis: a structural  
35 perspective of concept roles in semantic networks. *Comput.*  
36 *Soc. Netw.* **2**: 16.
- 37 Pavlopoulos, G. A., M. Secrier, C. N. Moschopoulos, T. G.  
38 Soldatos, S. Kossida, *et al.*, 2011 Using graph theory to an-  
39alyze biological networks. *BioData Min.* **4**: 10.
- 40 Pena, S. D. and R. Chakraborty, 1994 Paternity testing in the  
41 DNA era. *Trends Genet.* **10**: 204–209.
- 42 Pitrat, M. and C. Fauray, 2003 *Histoires de légumes: des origines*  
43 *à l'orée du XXI<sup>e</sup> siècle*. Institut National de La Recherche  
44 Agronomique (INRA), Paris, France.
- 45 Purugganan, M. D. and D. Q. Fuller, 2009 The nature of selection  
46 during plant domestication. *Nature* **457**: 843–848.
- 47 Sánchez-Sevilla, J. F., A. Horvath, M. A. Botella, A. Gaston,  
48 K. Folta, *et al.*, 2015 Diversity Arrays Technology (DArT)  
49 marker platforms for diversity analysis and linkage mapping  
50 in a complex crop, the octoploid cultivated strawberry (*Fra-*  
51 *garia* × *ananassa*). *PLoS One* **10**.
- 52 Sangiacomo, M. and J. Sullivan, 1994 Introgression of wild  
53 species into the cultivated strawberry using synthetic octo-  
54 ploids. *Theor. Appl. Genet.* **88**: 349–354.
- 55 Scott, J., 1988 Social network analysis. *Sociology* **22**: 109–127.
- 56 Shaw, P. D., M. Graham, J. Kennedy, I. Milne, and D. F. Mar-  
57 shall, 2014 Helium: visualization of large scale plant pedigrees.  
58 *BMC Bioinformatics* **15**: 259.
- 59 Simon, J. L. and P. Bruce, 1991 Resampling: A tool for everyday  
60 statistical work. *Chance* **4**: 22–32.
- 61 Sjulín, T. and A. Dale, 1987 Genetic diversity of North American  
62 strawberry cultivars. *J. Amer. Soc. Hort. Sci.* **112**: 375–385.
- Sjulín, T. M., 2006 Private strawberry breeders in California. *HortScience* **41**: 17–19.
- Staudt, G., 1988 The species of *Fragaria*, their taxonomy and  
geographical distribution. *Acta Hort.* **265**: 23–34.
- Staudt, G., 2003 *Les dessins d'Antoine Nicolas Duchesne pour son*  
*Histoire naturelle des fraisiers*. Publications scientifiques du  
Muséum, Paris, France.
- Telfer, E. J., G. T. Stovold, Y. Li, O. B. Silva-Junior, D. G. Grat-  
tapaglia, *et al.*, 2015 Parentage reconstruction in eucalyptus  
nitens using snps and microsatellite markers: a comparative  
analysis of marker data power and robustness. *PLoS One* **10**:  
e0130601.
- Thulasiraman, K. and M. Swamy, 1992 5.7 Acyclic directed  
graphs. In *Graphs: theory and algorithms*, p. 118, John Wiley &  
Sons, New York, NY.
- Trager, E. H., R. Khanna, A. Marrs, L. Siden, K. E. Branham,  
*et al.*, 2007 Madeline 2.0 PDE: a new program for local and  
web-based pedigree drawing. *Bioinformatics* **23**: 1854–1856.
- van Nocker, S. and S. E. Gardiner, 2014 Breeding better culti-  
vars, faster: applications of new technologies for the rapid  
deployment of superior horticultural tree crops. *Hortic. Res.*  
**1**: 14022.
- Vandeputte, M., 2012 An accurate formula to calculate exclusion  
power of marker sets in parentage assignment. *Genet. Sel.*  
*Evol.* **44**: 36.
- Vandeputte, M. and P. Haffray, 2014 Parentage assignment with  
genomic markers: a major advance for understanding and  
exploiting genetic variation of quantitative traits in farmed  
aquatic animals. *Front. Genet.* **5**: 432.
- VanRaden, P., 2008 Efficient methods to compute genomic pre-  
dictions. *J. Dairy Sci.* **91**: 4414–4423.
- Verma, S., N. Bassil, E. Van De Weg, R. Harrison, A. Monfort,  
*et al.*, 2016 Development and evaluation of the Axiom® IS-  
traw35 384HT array for the allo-octoploid cultivated straw-  
berry *Fragaria* × *ananassa*. *Acta Hort.* **1156**: 75–82.
- Voorrips, R. E., M. C. Bink, and W. E. van de Weg, 2012 Pedimap:  
software for the visualization of genetic and phenotypic data  
in pedigrees. *J. Heredity* **103**: 903–907.
- Wasserman, S. and K. Faust, 1994 *Social network analysis: Meth-*  
*ods and applications*, volume 8. Cambridge University Press,  
Cambridge, England.
- Whitaker, V. M., S. J. Knapp, M. A. Hardigan, P. P. Edger, J. P.  
Slovin, *et al.*, 2020 A roadmap for research in octoploid straw-  
berry. *Hortic. Res.* **7**: 1–17.
- Wickham, H., 2016 *ggplot2: elegant graphics for data analysis*.  
Springer-Verlag, New York, NY.
- Wilhelm, S. and J. E. Sagen, 1974 *A history of the strawberry, from*  
*ancient gardens to modern markets*. University of California,  
Division of Agricultural Sciences, Berkeley, CA.
- Williams, R. L., 2001 Bernard de Jussieu and the Petit Trianon. In  
*Botanophilia in Eighteenth-Century France*, pp. 31–44, Springer,  
Dordrecht, Netherlands.
- Wimmer, V., T. Albrecht, H.-J. Auinger, and C.-C. Schön, 2012  
synbreed: a framework for the analysis of genomic prediction  
data using R. *Bioinformatics* **28**: 2086–2087.
- Yu, H., P. M. Kim, E. Sprecher, V. Trifonov, and M. Gerstein, 2007  
The importance of bottlenecks in protein networks: correla-  
tion with gene essentiality and expression dynamics. *PLoS*  
*Comput. Biol.* **3**: e59.
- Zheng, X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, *et al.*,  
2012 A high-performance computing toolset for relatedness  
and principal component analysis of SNP data. *Bioinformatics*

