Supplementary Information S1 – Methods, Additional Table and Figures
Forest et al. - Preserving the evolutionary potential of floras in biodiversity hotspots

This document comprises the following items:

• Methods
• Supplementary Table 1 | Calibration points
• Supplementary Figure 1 | Study location and east/west subdivision
• Supplementary Figure 2 | Pairwise comparison between branch lengths generated under different phylogenetic optimality criteria
• Supplementary Figure 3 | Decoupled gains in genus richness and PD complementarity
• Supplementary Figure 4 | QDS with significantly low PD for their richness, as assessed under to measures of PD
• Supplementary Figure 5 | Frequency histograms of utility randomizations
• Supplementary Figure 6 | Taxon sets chosen to maximise PD significantly outperform choosing taxa at random at selecting useful genera.

Methods

Sampling and DNA sequencing. We sampled one exemplar species for 735 of the 943 genera of angiosperms currently recognized in the Cape (ca. 78%) and obtained sequence data for the plastid rbcL exon (ribulose-1,5-bisphosphate carboxylase/oxygenase large subunit). Voucher information and GenBank/EMBL accession numbers are provided (Supplementary Information S2). Of the 735 rbcL sequences included here, 395 sequences were produced for this study and 340 were downloaded from GenBank/EMBL sequence databases (211 of which are represented by species present in the Cape). DNA was isolated using a modified version of the 2× CTAB method1 and subsequently purified on cesium chloride/ethidium bromide gradients (1.55 g/ml density). PCR amplification of the rbcL exon was performed using primer combinations from Olmstead et al.2. PCR reactions (50 µl) were made with the ReddyMix PCR Master Mix from ABgene (2.5 mM MgCl2; Epsom, Surrey, UK) with the addition of 1 µl of bovine serum albumin 0.4 % and 50 ng of each primer. Program started with 2 min initial denaturation at 94°C, followed by 32 cycles of 1 min denaturation at 94°C, 1 min annealing at 48°C, 1.5 min extension at 72°C, and a final extension of 3 min at 72°C. After purification with the QIAquick kit (Qiagen, Inc.), cycle sequencing reactions were performed in 10 µl reactions using 1 µl of BigDye® Terminator cycle sequencing chemistry (v3.0 and v3.1; ABI; Warrington, Cheshire, UK) and run on ABI 377, ABI 3100, or ABI 3700 automated sequencers. SEQUENCHER 4.1 (Gene Codes Corp., Ann Arbor, Michigan, USA) was used to assemble complementary strands and verify software base-calling and sequences were aligned by eye in PAUP* (version 4.0b10; 3).

Phylogenetic analyses. Because of computing limitations imposed by the size of the matrix, phylogenetic relationships were reconstructed using the parsimony ratchet method4 as implemented in the software PAUPRat5 with 15% of the characters perturbed and 200 iterations. Ten independent parsimony ratchet searches were performed and the shortest trees resulting from these independent searches were used to create a consensus tree. For comparison purposes, a parsimony analysis was performed using the maximum parsimony criterion and heuristic search options as implemented in PAUP with 1,000 random addition replicates and tree bisection reconnection (TBR) branch swapping, retaining five trees per replicate to reduce time spent on swapping large numbers of suboptimal trees. The resulting trees were used as starting trees in a second analysis using the same parameters as above with a limit of 10,000 trees. These maximum parsimony searches did not recover shorter trees than the ratchet analyses (15,275 steps instead of 15,272 steps for the ratchet analysis); thus only results from the ratchet analysis are reported here. Clade support was assessed with the
bootstrap\textsuperscript{6} as implemented in PAUP* using 500 bootstrap replicates, subtree pruning-regrafting branch swapping, simple addition sequence and MulTrees ‘on’ but keeping only five trees per replicate\textsuperscript{7}. \textit{Amborella}, which is sister to the rest of the angiosperms\textsuperscript{8}, was used as outgroup.

**Dating and calibration.** One of the most parsimonious trees from the parsimony ratchet analysis was chosen as a best hypothesis of relationships for the Cape plant genera (Supplementary Information S3). PD calculations were performed using branch lengths (maximum parsimony and maximum likelihood) and absolute age estimates obtained using molecular dating techniques and calibration points from the fossil record. Maximum parsimony branch lengths were obtained using PAUP* and DELTRAN character optimization\textsuperscript{3}. Maximum likelihood branch lengths were optimized onto the tree using the HKY85+G+I model of DNA evolution\textsuperscript{9,10} with all parameters estimated from the data. This nucleotide substitution model was chosen because it represents a good compromise between model complexity and computing time. A likelihood ratio test\textsuperscript{11} showed significant rate heterogeneity across lineages (data not shown); therefore, age estimates were obtained using maximum likelihood branch lengths and non-parametric rate smoothing (NPRS) as implemented in the software \textit{r8s}\textsuperscript{12}. Relative time divergences were transformed into absolute ages using twelve well-characterized fossils as minimum constraints\textsuperscript{13,14} (see below; Table 1). A fixed age of 121 million years (Ma) was applied to the crown node of the eudicots based on the occurrence of tricolpate pollen grain in the fossil record, one of the most undisputed fossil dates for angiosperms\textsuperscript{15}.

**Distribution data.** The distribution of genera within the Cape was compiled as a binary matrix of absence/presence per quarter degree squares cells (QDS; approximately 25 km x 27 km) using data from the Pretoria National Herbarium database (PRECIS). QDS is the finest level of precision attainable with this database and the Cape comprises 201 QDS. Because we used generic distribution, we assume that the incidence of false absences in the data set is relatively low. Distribution data for some genera for which no information was present in PRECIS, were compiled from herbarium specimens housed at the Compton Herbarium (NBG) and from the literature. Some recent taxonomic rearrangements were not yet implemented in PRECIS; therefore, we modified distribution information according to these rearrangements. A single collection record was necessary to consider a genus present within a QDS.

**Phylogenetic diversity.** PD was calculated (and all other numerical and computational analyses save where noted) calculated in \textit{R}\textsuperscript{16} using the APE package\textsuperscript{17}. All three optimization methods examined here (MP; ML; absolute ages, NPRS) produced per-QDS PD values that are strongly correlated to each other (MP on ML, $R^2=0.9989$, $p <= 0.0001$; MP on NPRS, $R^2=0.9898$, $p <= 0.0001$; ML on NPRS, $R^2=0.9904$, $p <= 0.0001$); therefore we refer only to the results obtained with the NPRS age estimates.

To assess which QDS have a higher or lower number of genera in relation to their observed amount of PD, we calculated the residuals from a loess regression of the NPRS age estimates per QDS on per-QDS genus richness. To determine which QDS of the Cape have significantly higher or lower PD than that expected by chance alone, 10,000 randomizations were performed by calculating the PD for a random set of genera corresponding to the number of genera found in a given QDS. The observed total PD is compared to the distribution of the 10,000 random replicates (comparisons made at a significance level of 0.01). Randomizations were conducted using the NPRS absolute age estimates and repeated for each of the 201 QDS of the Cape.

**Phylogenetic correlation of medicinal and economic species.** A randomization procedure was used to assess if the distribution of medicinal and economic species is constrained by the phylogeny or randomly distributed across lineages. To be considered of
medicinal and/or economic use, a given genus must have at least one species found in the Cape that is recorded in the database of the Survey of Economic Plants for Arid and Semi-Arid Lands (SEPASAL\textsuperscript{18}), which comprises information for 6,200 species of known utility for humans. We scored the usefulness under three different categories: (1) medicinal, (2) food, and (3) all other uses except food and medicines (see Supplementary Information S2). Using the ‘shuffle’ option in the software MacClade\textsuperscript{19}, each binary character (presence/absence of use) was randomized among taxa (10,000 replicates) and the distribution of the frequencies of the resulting number of steps was used to test if the observed number of steps is significantly less than expected. Optimal taxon sets to maximise PD, as shown in Supplementary Information S1, Figure 6, were calculated using the program pda\textsuperscript{20}.

Supplementary Table 1 | Calibration points. Fixed and minimum constraints used to obtain node age estimates for the phylogenetic tree of the angiosperm genera of the Cape Floristic Region. Ages obtained from Schneider et al.¹ and Magallón and Sanderson² are given in million years. All ages were applied to the crown node of the given clades.

<table>
<thead>
<tr>
<th>Calibration (type)</th>
<th>Clades</th>
<th>Age</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (minimum constraint)</td>
<td>Poales</td>
<td>65</td>
<td>3, 4, 5</td>
</tr>
<tr>
<td>B (minimum constraint)</td>
<td>Laurales</td>
<td>105.5</td>
<td>6</td>
</tr>
<tr>
<td>C (fixed age)</td>
<td>Eudicots (tricolpate pollen)</td>
<td>121</td>
<td>7</td>
</tr>
<tr>
<td>D (minimum constraint)</td>
<td>Caryophyllales</td>
<td>83.5</td>
<td>8</td>
</tr>
<tr>
<td>E (minimum constraint)</td>
<td>Saxifragales</td>
<td>89.0</td>
<td>9</td>
</tr>
<tr>
<td>F (minimum constraint)</td>
<td>Myrtales</td>
<td>85.8</td>
<td>10</td>
</tr>
<tr>
<td>G (minimum constraint)</td>
<td>Malvales</td>
<td>68</td>
<td>11</td>
</tr>
<tr>
<td>H (minimum constraint)</td>
<td>Cucurbitales</td>
<td>54.8</td>
<td>8</td>
</tr>
<tr>
<td>I (minimum constraint)</td>
<td>Solanales</td>
<td>33.7</td>
<td>8</td>
</tr>
<tr>
<td>J (minimum constraint)</td>
<td>Ericales</td>
<td>89</td>
<td>12</td>
</tr>
<tr>
<td>K (minimum constraint)</td>
<td>Gentianales</td>
<td>33.7</td>
<td>13, 14, 15</td>
</tr>
<tr>
<td>L (minimum constraint)</td>
<td>Apiales</td>
<td>37</td>
<td>16</td>
</tr>
</tbody>
</table>

Supplementary Figure 1 | Study location and east/west subdivision. The east-west divide corresponds to the boundary between the winter-rainfall (shaded) and mainly non-seasonal rainfall (open) zones of the Cape. The PD/taxon aggregated across all QDS is ca. 1.7% higher in the eastern part of the Cape: $2.42 \times 10^7$ year/taxon compared to $2.38 \times 10^7$ years/taxon in the western part, which translates to 70.8 million years for a QDS of average genus richness.

Topological base map is from the UNEP World Database on Protected Areas (http://www.unep-wcmc.org/wdpa/)

Supplementary Figure 2 | Pairwise comparison between per-QDS PD generated under different phylogenetic optimality criteria. Regressions (upper triangle) are standard linear (red) and loess (green) with a span (\(\propto\)) of 0.75 x the range of the predictor variable. Adjusted R\(^2\) values for the linear regressions are given in the lower triangle. The branch lengths for the taxonomic PD measure are in arbitrary units, with one unit being allocated to differences between each of the 8 levels of the APG taxonomy (genus, family, order, and 5 unnamed higher taxa).
Supplementary Figure 3 | Decoupled gains in genus richness and PD complementarity. Solid lines indicate performance when QDS sets are chosen based upon either genus richness (blue) or PD (red, branch lengths proportional to evolutionary time under NPRS). Dot-dash lines indicate performance when the alternative optimality criterion is chosen (gains in genus richness when PD is optimised and vice versa). Grey denotes the 5th and 95th percentile of 1000 random QDS sets of the relevant size. Note that, as would be expected, the early cell selections under all optimality criteria are much more efficient than the randomisations, leading to a significantly more rapid capture of PD and richness.
Supplementary Figure 4 | QDS with significantly low PD for their richness, as assessed under two measures of PD. The lower figure shows cells with significantly low “taxonomic PD” as assessed against the APG taxonomy (see legend, Fig.2) for their genus richness. The upper figure, showing cells with significantly low NPRS PD is taken from the main manuscript for comparison.
Supplementary Figure 5 | Frequency histograms of utility randomizations. Results of the randomization procedure (10,000 replicates) used to assess if the distribution of genera with useful species is correlated with the phylogeny. For all three categories of uses (food, medicinal and others uses) considered separately (a, b, c), the randomizations show that usefulness is significantly correlated to the phylogeny (p<0.01). Arrows point to the observed number of steps (food uses, 69 steps; medicinal uses, 52 steps; other uses, 89 steps).

a. Food uses

b. Medicinal uses

c. Other uses
Supplementary Figure 6 | Taxon sets chosen to maximise PD significantly outperform choosing taxa at random at selecting useful genera. The genera chosen to maximise PD (red points) contain a higher number of useful genera (all classes of use combined) than do the same number of genera selected at random (empirical 0.95 quantile from 25000 randomisations – grey; median - black) for sample sizes indicated by the red dots at y=0. Note that the first partial cell set in Fig. 2 (a single QDS) contains 479 genera. As such, the conservation scenarios in Figure 2, which show decoupled gains in PD and taxon complimentarity, do not show decoupled gains in useful genera under the two optimality criteria. This suggests that the relative benefits of PD maximisation may depend in part upon the scale of the phylogenetic distribution of the precise features of genera found to be useful in the future. We note, however, that cells with high PD for their genus richness (residuals, Fig. 1d) show a slightly but significantly higher mean proportion of useful genera than do cells with low PD for their genus richness (0.260 vs 0.239, p<0.01, t-test of differences in mean arcsine-transformed proportions). Optimal sets of n taxa to maximize PD were found using the program pda20.