Pangenome analysis identifies gene loss in Scalesia across the Galápagos Islands.

**Abstract:**

We present the first genus-level pangenome of *Scalesia* – Darwin’s giant daisies. This pangenome is composed of thirty-four individuals representing eighteen *Scalesia* species, collected among ten islands of the Galápagos. Gene presence-absence variation analysis identified 1998 variable genes amongst a total of 43093. Gene ontology enrichment identified Germacrene A derived terpenoid biosynthesis as well as other secondary metabolic processes amongst the variable gene set. Absence of genes involved in these processes may contribute to the formation of the “plant island syndrome” phenotype displayed in *Scalesia* species. Gene presence-absence profiles also differ between islands. We suggest that this pattern supports the current theory of adaptive radiation of *Scalesia* across the archipelago. This work implies gene-absence profiles between sub-populations can be applied to support migratory theories, in relation to the adaptive radiation of plant island species.

[A few more sentences on maker results once complete].

**Introduction:**

*Scalesia*, or Darwin’s giant daisies, are a genus of the Asteraceae family, endemic to islands of the Galápagos. Unlike their smaller mainland ancestors, species of *Scalesia* present as large shrubs, with some species growing into full, forest-forming trees. The genus has long been recognised as an exemplary example of adaptive radiation, diversifying rapidly to occupy various ecological niches among the archipelago, with some species localised to a single island. *Scalesia* demonstrate remarkable ecological and morphological diversity, occupying habitats ranging from arid lowlands to humid highland forests (1). This diversity, shaped by the unique environmental gradients of the islands, has led to *Scalesia* being an important model for investigating the genomic mechanisms underlying plant adaptation and speciation in insular environments (2).

Since Darwin recognised the intriguing variation presented by *Scalesia* in 1839 (3), modern evolutionary biologists have leveraged the genus to gain insights into the effects of invasive species (4) (5) and as a model to investigate the phenomenon of adaptive radiation (6). Fernández-Mazuecos *et al.* (2020) presented a phylogenomic analysis of 15 *Scalesia* species, estimating time passed since divergence from its South American mainland relatives and subsequent diversification throughout the archipelago. They suggested an initial *Scalesia* common ancestor colonised the younger, central islands, of the Galapagos before radiating out to the older island of San Cristóbal. Rapid intra-island speciation then followed to give rise to *Scalesia* species recognised today. This pattern of migration is counter to the “progression rule”, suggesting older lineages inhabit the oldest islands, as has been observed in other plant island species (7) (8) (9). Recently, molecular insights into the genus have enhanced, with the publication of a reference genome for the critically endangered *Scalesia atractyloides* (10). This work provided valuable advances into understanding the genetic profile of the *Scalesia* genus common ancestor, providing evidence it was an allotetraploid. However, single-reference genomes are inherently limited in their ability to represent genome-wide variation within a linage. They fail to capture presence-absence variation (PAV), gene duplication and structural variants, elements that may be important for adaption and diversification. These limitations are amplified in species produced through adaptive radiation, in which evolutionary novelty may arise through lineage-specific gene content.

Pangenomics has emerged to overcome the limitations of single reference genomes by aiming to capture all genomic content within a population. This includes core genetic material shared by all individuals, as well as accessory or variable material only present in a subset (11). The value of pangenomic analysis has been demonstrated in species such as *Brassica oleracea*, in which gene PAV has been linked to agronomic traits (12). In soybean and tomato, pangenome studies identified novel genes absent from their reference genomes and are associated with disease resistance and environmental adaption (13) (14). These case-studies illustrate how pangenomes can identify functionally and evolutionary important genomic regions, overlooked by traditional, single genome references. Pangenomes enable the identification of gene conservation and loss, that has been recognised as a contributing factor for diversification of species (15). However, the role and importance of gene PAV driving adaptive radiation of island plant species remains understudied. Island species often face strong ecological pressures, including drought, nutrient limitations and isolation that over time, shape their genomic architecture (10) (16). This observation has led to the formulation of the “plant-island syndrome” hypothesis, in which species undergo convergent evolution, manifesting in traits including increased size, woodiness and reduced defence against herbivores (17). Such convergent traits allow plant island species to streamline their ability to survive in specialised island niches. *Scalesia* presents these traits (10), however, whether and to what extent, gene conservation and loss has had in driving the formation of these phenotypes has not yet been documented.

Here we constructed a genus-level pangenome representing 18 accessions of *Scalesia*, employing an iterative mapping approach (18) to gain an understanding of how genomic PAV may contribute to *Scalesia*’s adaptive radiation across the Galápagos. This study presents the first observations of gene conservation and loss within *Scalesia* and suggests that this may have contributed to the evolution of traits typically associated with plant-island syndrome, primarily by specialisation of *Scalesia’s* secondary metabolism. Notably, this includes downregulation of Germacrene A derived terpenoids, molecules known to be involved in defence against herbivores (19,20). Additionally, we assessed whether differences in gene content between *Scalesia* individuals can be used to support the current migratory theory of the genus across the Galápagos (6). Our results indicate PAV profiles of individuals from the same or neighbouring islands tend to be similar compared to geographically distant individuals, as indicated by UMAP analysis. We show that gene loss in *Scalesia* increases with distance from the central islands of the Galápagos, a pattern that supports the current theory of *Scalesia* migration by adaptive radiation across the archipelago. These findings enhance the current understanding of the genomic mechanisms underlying adaptive radiation in island plants. Beyond its evolutionary importance, insight into gene loss of *Scalesia* species may inform conservation strategies and preserve biodiversity within the Galápagos.

**Results**

**General MAKER Pangenome results.**

[To complete]

**Spatially Close Individuals Share Similar Feature Absence Profiles.**

To investigate gene presence-absence variation in *Scalesia*, we collated genome sequences for 34 *Scalesia* individuals collected from ten different islands around the Galápagos Archipelago (Figure 1). This dataset represents eighteen different *Scalesia* species including two *Scalesia* hybrids. PAV analysis identified 1998 variable genes across the *Scalesia* individuals from a total of 43093 total genes (4.64%).

We employed UMAP clustering to assess the extent and pattern of gene absence variation between individuals and observed that individuals tended to cluster with other individuals from the same island or neighbouring islands (Figure 2). For example, all three individuals from Isabela clustered together, geographically furthest in distance from individuals obtained from San Cristóbal, located at the eastern end of the archipelago. UMAP clustering also revealed that individuals from Santa Cruz formed a distinct cluster, separate from individuals from San Cristóbal and the westerly islands of Santiago and Isabela. Despite forming general clusters, there were several outliers. For example, the *S. pedunculata* individual from Santiago had a PAV profile more representative of individuals from San Cristóbal, rather than those from the same or the neighbouring island Bartolomé. Samples obtained from the Southernmost Island of Floreana presented split clustering, with one sample clustering amongst samples from San Cristóbal, and two samples more representative of individuals obtained from the central islands – Isabela, Santiago/ Bartolomé and Santa Fe.

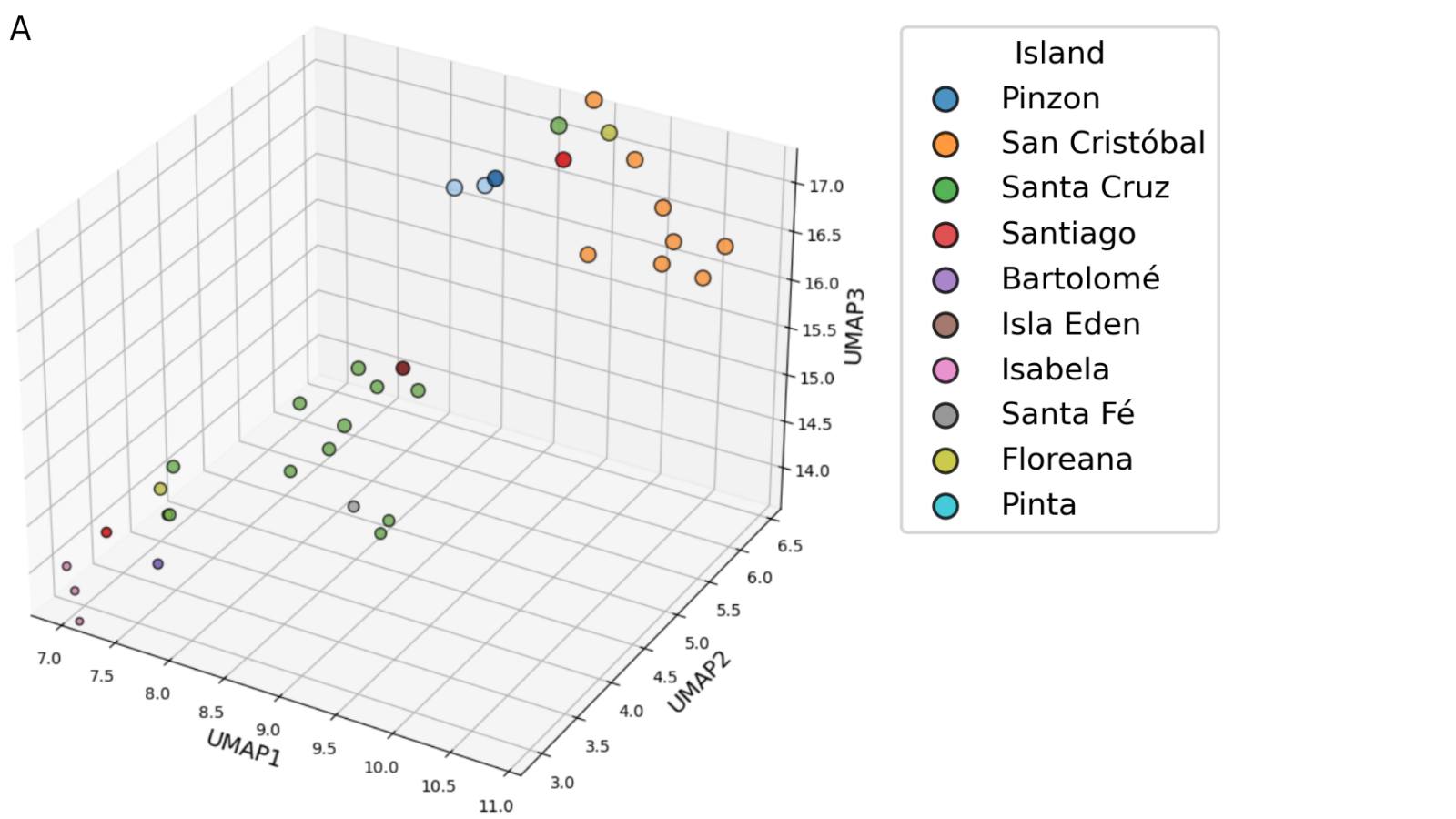
**Figure 1.**

A map of islands with names

AI-generated content may be incorrect.

**Figure 1. Map of the Galápagos Islands with longitude/latitude coordinates of individuals obtained for the study.** Colour indicates which island the individual was taken from. In the case of Bartolome and Isla Eden, these are small islands off the coast of Santiago and Santa Cruz respectively. These islands were too small to add to the island overlay; their relative positions can be visualised using each individual’s co-ordinates.

**Figure 2.**



**Figure 2**. **UMAP clustering reveals similarity between accessions from the same or geographically similar islands.** 3D UMAP generated using Hamming’s distance on the PAV matrix (n=34). UMAP was run with parameters n\_neighbours=7 and random\_state=42. Model was run using UMAP Python package. Point size scaled using UMAP-Z co-ordinate to enhance depth perception. Points coloured according to island.

**Gene Ontology Enrichment Analysis Indicates Absence of Genes Relating to Secondary Metabolic Pathways.**

Gene ontology (GO) enrichment analysis of the annotated variable gene set (n=1508) identified 45 significantly enriched GO terms (P <0.01). Figure 3.A shows all 45 GO terms in word cloud format, with all FDR values accessible in Sup. Table 2. The top 15 most significantly enriched GO terms for each GO namespace are shown in Figure 3. B. These enrichment results indicated a general loss of genes involved in secondary metabolite biosynthesis. More specifically, the study highlighted multiple GO biological processes associated with terpenoid/isoprenoid metabolism. Terpenoid production for defence against herbivores in plants has been well studied (19,20) and enrichment of such genes within the variable subset implies *Scalesia* species may be downregulating production of these compounds via gene loss.

**Figure 3.**

A

A close-up of text

AI-generated content may be incorrect.

**A close-up of a graph

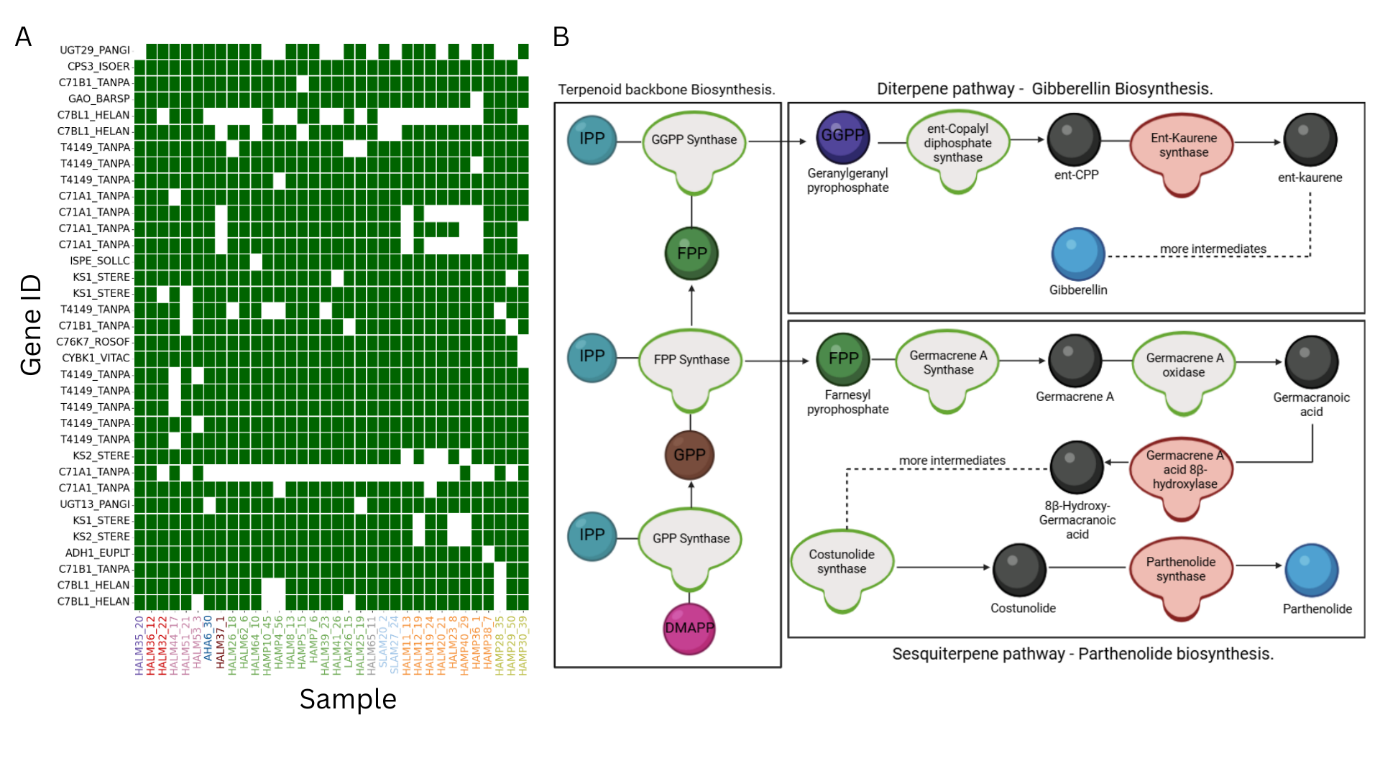
AI-generated content may be incorrect.**

**Figure 3**. **Gene ontology (GO) analysis reveals enrichment of secondary metabolic processes within variable gene group. (A).** Word cloud containing all 45 significantly enriched GO terms (P <0.01). Figure was generated using the Python Wordcloud package. Font size of scaled by p-value. **(B).** Top 15 (maximum) GO terms grouped by GO namespace, sorted by FDR value. Dashed black line indicates equivalent p-value of 0.05.

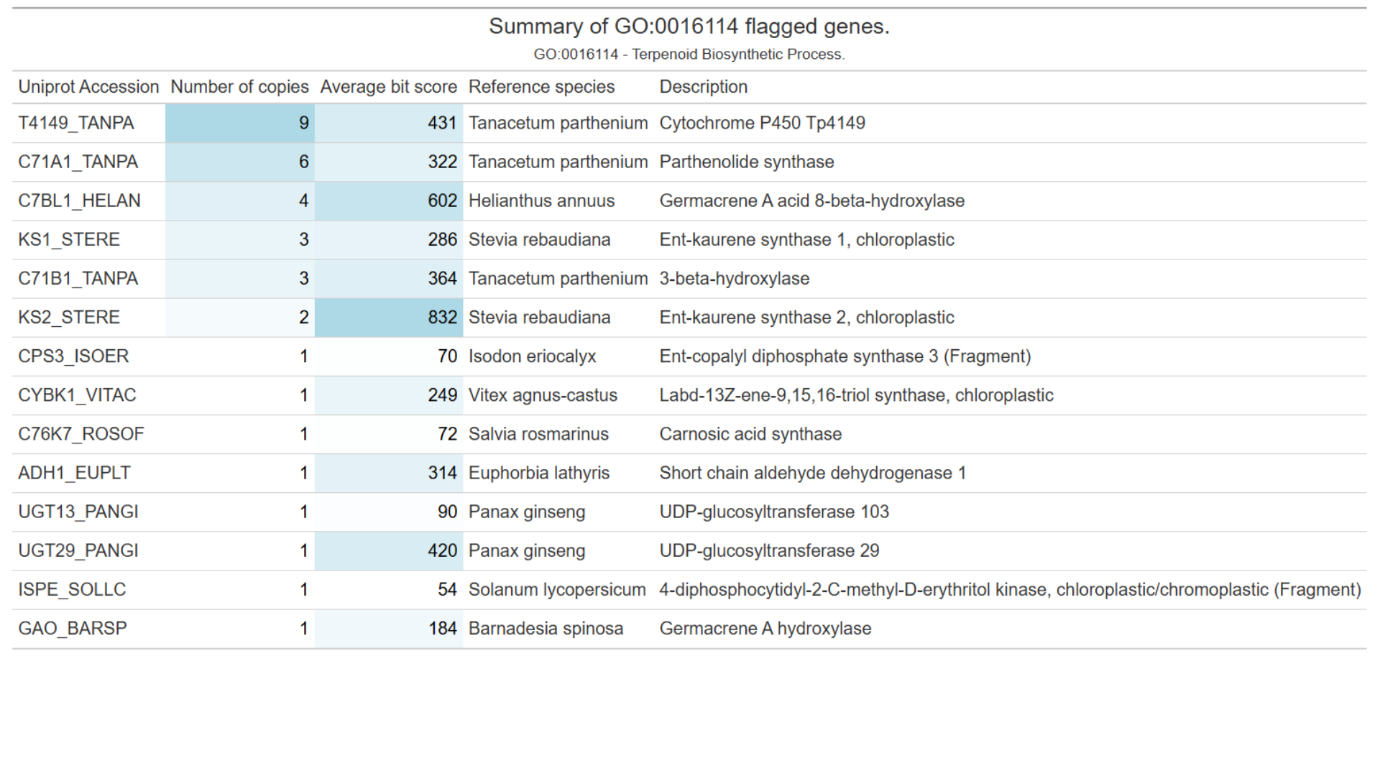
To gain more insight into the specific genes responsible for the enrichment of these GO terms, we plotted presence/absence for each of the variable genes associated with the top GO biological processes as outlined in Figure 3.B and sorted them by genomic location based on position within the linearised reference genome. (Figure 4.A). Amongst these genes was Parthenolide synthase (C71A1\_TANPA), returning high-confidence hits to six different variable genes. This enzyme regulates the conversion of costunolide to parthenolide, responsible for the final step in parthenolide biosynthesis from Farnesyl pyrophosphate (FPP) (45). Parthenolide is a sesquiterpene lactone – a subset of terpenoids predominantly found within the Asteraceae family (46). Studies linking sesquiterpene lactones to herbivore deterrence began as far back as 1982 (47). Figure 4.A indicates many individuals are missing Parthenolide synthase at one locus, with a subset of individuals primarily obtained from San Cristobal (orange), showing similar absence of Parthenolide synthase at three loci clustered together. Except for the individual from Isla Eden, all other individuals were flagged as present for these loci.

Germacrene A acid 8-beta-hydroxylase is an upstream member of the same pathway, responsible for the intermediary conversion of germacrene A acid to 8-beta-hydroxy-germacrene A acid (C7BL1\_HELAN)(48). This enzyme returned high-confidence hits to four different loci within the variable gene set. Visualisation of the presence/absence pattern of these genes indicated varying degrees of absence between individuals. One locus indicated heavy absence across individuals from San Cristobal, while absence was less common among individuals from the central islands. Individuals from Isabela, Bartolomé and the *Scalesia stewartia* from Santiago all retained presence at this locus. Cytochrome P450 Tp4149 (T4149\_TANPA) presented high-confidence hits to nine different loci within the variable gene set (Table 1). It has been suggested that this enzyme also plays a role in the same biochemical pathway, however its exact function remains undetermined (45). Figure 4.A indicated that these loci clustered at two regions within the genome, at one of these regions, including 5/9 gene copies, absence of these features only occurred in two individuals from Isabela, with remaining individuals showing uniform presence.

We also identified gene absence variation in Ent-kaurene synthases 1 and 2 (KS1\_STERE & KS2\_STERE), at three and two loci respectively. These enzymes have not been shown to be involved in production of sesquiterpene lactones and instead are responsible for diterpenoid gibberellin biosynthesis – a separate branch of terpenoids (49). Rather than defence response, diterpenoid gibberellins function as plant growth regulators derived from a distinct branch of the core terpenoid backbone pathway (50). An overview diterpenoid and sesquiterpene terpenoids from the common terpenoid backbone pathway is outlined in Figure 4.B, enzymes belonging to the variable gene set are highlighted red. Gene descriptors for each gene are shown in Table. 1.

**Figure 4. Gene enrichment analysis reveals variable gene content of terpenoid pathway through multiple enzymes. (A).** Visual representation of the PAV matrix for all variable genes associated with GO term GO:0016114: Terpenoid biosynthetic process. Accession (x axis) is sorted by Island relating to colour key present in Fig.1 and 2. Green indicates presence, white absence. **(B)**. Core enzymes and biochemical pathways responsible for gibberellin and Sesquiterpene production. Enzymes marked as non-core are highlighted in red. Figure generated using BioRender (ref).

**Table 1.** Table representing all unique genes within variable gene set mapping to GO term 0016114 – Terpenoid biosynthesis. Number of copies represents the number of times the gene appears in the non-core dataset. Average bit score represents the mean bit-score as obtained from `blastp`search- if gene is mapped to multiple non-core loci. Description obtained from “stitle” column during `blastp` mapping.



[Could also look at the specific absent genes relating to lipid and hydrocarbon metabolism – or for discussion as evidence of specialising metabolism in relation to PIS.]

**Spatial Variation in Gene Presence–Absence Reflects Progressive Gene Loss Away from Galápagos Central Islands.**  
We calculated distance in longitude/latitude of each individual from the centre of Santiago Island and looked for associations with presence-absence data. We first tried to identify trends using the entire functionally annotated PAV matrix (n = 1508) and identified no significant relationship between the two variables Sup4.A. We filtered the annotated variable gene set to retain only those mapping to any of the top ten most enriched GO terms under the “biological process” namespace (GO: BP, n = 91). We identified a significant negative correlation between distance from Santiago and number of present genes using two separate regression models (OLS P < 0.001 and Poisson P = 0.037).

**Figure 4**.

A graph with many colored dots

AI-generated content may be incorrect.A diagram of a graph

AI-generated content may be incorrect.

**Figure 4. Gene loss in GO enriched processes increase with distance from Santiago Island. (a).** Scatter plot representing number of genes marked as present (1) within the PAV matrix, subset to retain only genes mapped to at least one of the most significantly enriched GO terms under namespace “biological process”, against geodesic distance from Santiago. Individuals are coloured by island using key in Figure 1. A Dashed best-fit linear regression line across all points with 95% confidence intervals was overlaid to show trend. **(b).** Neighbour-joining tree calculated using Hamming’s distance on the Variable PAV matrix from (a). Individuals coloured according to key.

[Other ideas]? – Maybe we need to run the same analysis from the other central islands as well as Santiago, to see which provides the biggest negative effect? – Could build a panel to show there is no significance from any other island?

**Discussion.**

**Maker Results Paragraph:**

* Does the number of core and variable genes reflect that of other plants?
* High number of gene repeats implications.
* High level of contamination in individuals.

Presence-absence analysis of 34 *Scalesia* individuals identified 1998 variable genes that may contribute to the diversification of different *Scalesia* species. Initial UMAP analysis indicated a clustering of individuals according to island. However, some individuals exhibited PAV profiles more characteristic of those in geographically distant islands. For example, the *S.* *pedunculata* individual from Santiago, clustered with individuals from San Cristóbal, rather than with geographically nearer individuals of Santiago and Bartolomé. This pattern is maintained in the neighbour joining tree (Fig 4.B), also showing the *S. pedunculata* individual from Santiago groups with individuals from San Cristóbal, rather than other individuals from Santiago or neighbouring islands. In addition, the *S. pedunculata* individual from Santa Cruz also revealed a comparatively different PAV profile compared with other individuals obtained from Santa Cruz, clustering close to *S.* *pedunculata* from Santiago in both neighbourhood-joining tree and UMAP analysis. Phylogenetic data presented by Fernández-Mazuecos *et al. (2020),* confirmed that S. *pedunculata* is a late diverging species within the *Scalesia* genus, more closely related to species on San Cristóbal than species such as the comparatively early diverging *S. stewartii* of Santiago. A possible explanation for these observations may the occurrence of back-population among islands, specifically from San Cristóbal to Santiago and Santa Cruz, however more research is required to confirm this hypothesis.

GO enrichment analysis of functionally annotated variable loci identified significant enrichment of GO terms associated with terpenoid biosynthesis and other secondary metabolic functions. Downregulation of anti-herbivorous terpenoid compounds has previously been documented in island-plant species(19). More specifically, parthenolide - a well-studied sesquiterpene lactone, has been shown to act as an anti-herbivorous agent in feverfew (*Tanacetum parthenium*), a fellow member of the Asteraceae family (51). The relatively few large herbivores within the Galapagos, compared with the mainland (52) may lead to a reduction in the production of herbivore defensive compounds.

The observation that soil quality and mineral availability varies considerably within the Galapagos also supports this finding (53). Interestingly, Strahlhofer *et al.* (2021) observed the older island San Cristóbal had noticeably poorer soil fertility than the younger islands of Isabela and Santa Cruz, whilst we highlighted increased gene loss on San Cristobal in this study. This phenomenon may have contributed to *Scalesia* streamlining metabolism through gene loss, downregulating expression of metabolically expensive secondary compounds such as terpenoids to adapt to less fertile soils. Further studies are needed to investigate if there is any causal relationship between soil fertility and gene loss.

Loss of Parthenolide synthase, Germacrene A acid 8-beta-hydroxylase and potentially Cytochrome P450 Tp4149 suggest a downregulation of terpenoid synthesis, though as some of these genes are present in multiple copies and some copies remain, it is unlikely terpenoid synthesis has been lost completely. It is more probable gene loss leads to a diversity in the levels of sesquiterpene lactone production across *Scalesia* species of the Galápagos, rather than loss of the process altogether. The observation of gene loss in ent-kaurene synthases 1 and 2, involved in gibberellin biosynthesis was surprising, as Darwin’s Daisies are a textbook example of island gigantism in plant species. The biological functions of gibberellins are well studied, being widely accepted that these compounds stimulate growth (50,54). Further studies are required to identify why ent-kaurene synthase 1 and 2 genes are undergoing gene loss. This data may hint at secondary functions for these enzymes. Alternatively, it is also possible this gene loss could be an adaptation to less fertile soil as previously discussed. The interaction between branches of the terpenoid pathway (Figure 4.B) is also interesting. As both branches of the pathway share the same substrate, downregulation of Germecrene-A based terpenoids has potential to free up substrate, allowing for upregulation of diterpenes ultimately leading to gibberellin synthesis that may increase plant growth. This hypothetical interplay between terpenoid branches may affect the formation of the island gigantism observed in *Scalesia* across the archipelago.

Reduction in biosynthesis of defensive compounds is a generally considered a trait of “plant-island syndrome”, as has been observed that threat from large-herbivores and pathogens pressures tend to be lower in insular environments when compared to mainland populations.(17). However, recent meta-studies assessing the prevalence of anti-herbivorous traits in island species have produced mixed and even contradictory results, suggesting the relationship may be more complicated than previously assumed. (31). Factors such as island age, degree of isolation and historical presence of herbivores are likely to influence the degree to which plant-island species converge on reduced herbivore defence. Due to these factors, further molecular-based studies are required to confirm whether *Scalesia* does indeed present this phenotype.

In line with previous evidence, we hypothesised the pattern of feature absence observed may be associated with the evolutionary migration pattern of *Scalesia* across the Galápagos. Using prior work by Fernández-Mazuecos *et al. (2020)*, suggesting a common ancestor of *Scalesia* landed and migrated outwards from the central islands of the archipelago (Santiago, Santa Cruz, Floreana, and Isabela), before colonising the older island of San Cristóbal. We attempted to determine whether gene loss away from the central islands followed a similar pattern. When we assessed this association in the variable gene set, we did not identify a significant trend SF4.A. Sub-setting the variable PAV matrix to retain only genes mapped to at least one of the top 10 most enriched GO biological processes revealed significant a negative correlation in multiple models Fig 4.A. Exactly why this subset caused exaggeration of the trend, and its subsequent significance, remains to be determined. An explanation for this observation may be that GO term filtering reduced the impact of genes incorrectly flagged as absent within the dataset, as GO term enrichment arises due to the absence of multiple genes. Despite the significant correlation that we observed, it is obvious that the number of present genes varies considerably between individuals of the same island Sup 4.B. For example, in individuals from San Cristóbal, the number of present genes varied between 1220 and 1310 min/max respectively, resulting in a range of 90 PAVs. This observation, when paired with factors such back-migration, can introduce substantial amounts of noise within the dataset, which may obscure true trends. Hard conclusions based solely on PAV data should be interpreted with caution and instead should be used as supporting evidence for pre-existing findings, or in combination with other phylogenomic approaches to concrete novel theories.

[There may be more to discuss here if we expand on geographical and environmental gene loss findings]. - Mention that we were unable to find any associations between number of present genes and environment climate/habitat? No point doing by species as most only have 1 sample.

These results suggest gene loss within *Scalesia* has likely contributed to the formation of the “plant-island-syndrome” phenotype. We identified 1988 previously unidentified variable regions of the *Scalesia* pangenome and subsequent analysis of these regions uncovered evidence of loss of multiple genes involved in Germacrene A derived biosynthesis. Using these variable regions, we highlighted PAV profiles of *Scalesia* individuals tend to vary less among geographically closer individuals and that this trend can be used as supporting evidence for the current migratory theory of *Scalesia* as proposed by Fernández-Mazuecos *et al.* (2020). This study would benefit from true population defining workflows such as STRUCTURE (55) to highlight inter-island populations. In turn, this may have enabled cleaner trends to emerge, and to help counteract the limited individual size of this dataset. A larger sample size is required to more accurately examine how island, accession and habitat type are associated with *Scalesia* gene loss in the Galapagos. We did not have any individuals corresponding to the reference genome *Scalesia atractyloides*, which may have led to an underestimation of core genes and an overestimation of accessory genes. In summary, this study demonstrates pangenome analysis can provide novel insights into the manifestation of plant-island syndrome, and how gene loss may contribute to the formation of new species, particularly in the context of adaptive radiation.

**Materials and Methods:**

**Pangenome Assembly.**

This protocol has been used to generate pangenomes for multiple plant species in the past (22,23). Illumina paired-end sequencing reads were obtained for 34 *Scalesia* individuals collected from the Galápagos Islands, collectively representing 18 different *Scalesia* accessions (Sup. table 1). Initial read quality metrics were obtained using FastQC (v0.11.9). Paired-end reads underwent adapter trimming and quality control using Trimmomatic (v0.11.9) (24).

Post-trimmed quality was assessed with FastQC/MultiQC (v1.12) (25). Trimmed reads were mapped to the pre-published *Scalesia atractyloides* reference genome (10) using Bowtie2 (v2.4.5) (21). Sorted BAM files for the 34 individuals were generated with Samtools (v0.5.0) (26) using *`*view*`*, *`*sort*`* and `index*`* commands with default parameters. Initial alignment metrics were generated via Samtools *`*stats*`* and visualised using a custom Python script [GitHub].Alignment metrics indicated a high percentage mapping rate, with a minimum alignment percentage of 93.14% within the sample set (Sup. Figure 1).

The primary pangenome was generated by extracting unmapped reads from each BAM file using Samtools (-f 4) flag. Unmapped read headers were used to retrieve their relative uncleaned counterparts, which were assembled *de novo* with MaSuRCA (v4.14) (27). BBmap (v39.33) (28) was used to estimate average insert size and standard deviation of individuals (231, 33.2 respectively). These values were passed to MaSuRCA as input parameters, as well as JF\_SIZE=200000000 and cgwErrorRate=0.15. The resulting *de novo* assembly contained 171908 contigs. This assembly was filtered using BLAST (v2.16.0) (29) to remove organelle contaminants against a custom dataset built using the *Helianthus annuus* NCBI accessions: NC\_007977.1 (chloroplast) and NC\_023337.1 (mitochondria) genomes. Blast hits with an e-value < 1e-10 were removed from further processing. The filtered assembly was mapped back to the original *Scalesia* reference using Minimap2 (v2.30) with default parameters (30) .Unmapped contigs (57783 contigs) were extracted with Samtools (f -4) and queried against the core-nt database with Blastn search. Taxonomy information was annotated to blastn hits with Taxonkit (v0.20.0) (31). Contigs matching “streptophyta”, the green plants kingdom, and “novel” contigs returning no blastp hits were retained. This file, containing 2352603 base pairs across 6338 contigs was concatenated to the *Scalesia atractyloides* reference genome to build the final pangenome. Hits per species and kingdom were visualised with Python Plotly (SF3).

**[ Add in MAKER details for final pangenome annotation].**

**Calling Presence/Absence Variation.**

The GFF annotation file for the reference dataset was downloaded via the DOI above and utilised in tandem with SGSGeneLoss (v0.1) (32) to calculate presence-absence variation of each individual against the reference dataset on the 34 *Scalesia* BAM files. SGSGeneLoss was run using a minimum coverage (mincov) value of 0.05, this value has become standardised across other gene-loss studies (33,34). We generated presence-absence calls of the 34 individuals across a total of 43093 annotated features within the GFF file. A custom Python script [GitHub] was developed to parse SGSGeneLoss output `.excov` files to determine if each individual contained adequate read depth/coverage to prevent the calling of false absence of features (Sup. Fig2). All 34 individuals were considered suitable for further analysis based on read depth and number of features marked present. A custom Python script [GitHub] was used to build the presence/absence data for each individual, before merging with Pandas (v2.2.3) (35) to generate the binary PAV binary matrix. Variable features – defined as features absent in at least one individual were filtered for within the same Python script.

Output files from the SGSGeneLoss package were also parsed to assess whether sequencing depth was sufficiently high (~10x coverage or higher to account for real absence of genetic features) (S2. B). The PAV matrix was filtered to identify 1988 unique features within the dataset hereby defined as variable (missing in at least one individual). variable

**UMAP clustering of PAV matrix.**  
UMAP analysis was run using the Python package UMAP (v 0.5.9) (36). The model was run using parameters: random\_state=42, metric="hamming" and n\_neighbours=7. All other parameters were left as default values. UMAP models were run using n\_components=2 and 3 to generate 2D and 3D plots respectively. For the 3D UMAP, points were scaled using the UMAP-Z coordinate to enhance depth perception, legends were removed for clarity. UMAP analysis was performed on the PAV matrix and points were coloured using individual metadata (Sup. Table 1).

**Functional Annotation of The *Scalesia atractyloides* Reference GFF.**  
Protein sequences were extracted for each feature of the *Scalesia atractyloides* GFF annotation file using the Gffread package (v0.12.7) (37) with default parameters. The resultant protein sequences were cleaned using a custom Python script [GitHub] to ensure all protein sequences were compatible for BlastP search. The Diamond package (v2.1.2) *`*makedb*`* command was used to generate a Diamond database file from the Uniprot/Swiss Prot collection of annotated proteins by following the instructions available here: <https://github.com/bbuchfink/diamond/wiki/3> . Under section :*`*Makedb Options*`*, run with flags --taxonmap, --taxonnodes and –taxonnames (38) (39). The Diamond database was used to run a `blastp` search of all GFF file proteins.Hits with e-value < 1e-5 were considered statistically significant. Resulting BlastP hits were annotated with GO terms using the publicly available Go term mapping file available here: <https://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/idmapping/idmapping_selected.tab.gz>, using a custom Python script and the Obonet package (v1.1.1) to merge GO terms with BlastP hits using the 'staxids' field (40). The top high confidence hit (if multiple hits) was used to provide functional annotation of the feature.

**Gene Enrichment Analysis.**  
A custom Python script [GitHub] was generated to map GO terms from BlastP hits to human-readable descriptions using the publicly available mapping file *`*go-basic.obo*`* available for download here: <https://geneontology.org/docs/download-ontology/> .  
Go enrichment analysis was performed using the *`*Goatools*`* (v1.4.12) (41) Python package using the function GOEnrichmentStudy() with parameters: `alpha=0.01, methods=["fdr\_bh] propagate\_counts=True. Fishers exact test was used to assess significance. GO terms with P < 0.01 were considered significant.

**Geographic Gene Loss Association.**

Geographic distances between each individual and Santiago Island (-0.252°N, -90.718°W) were calculated using geodesic distance calculations to account for curvature using GeoPandas (v 1.0.1) (42). This Santiago was chosen as the geographic centre for all distance-based analyses.

We assessed association of presence-absence and geographic distance from Santiago within the variable gene set by running Ordinary Least Squares (OLS) and Poisson regression analysis using the statsmodels (v 0.14.5) (43) Python package. Neither model identified a significant relationship between distance from Santiago and the number of absent genes (SF4.1). The variable gene PAV matrix was filtered to identify genes that mapped to the top 10 most significant GO terms flagged in the GO enrichment study based on FDR value. The total number of present genes for this subset were calculated for each individual and plotted to visualise association with distance from Santiago Island (kilometres) using matplotlib (v3.10.0) (44). We reran OLS and Poisson statistical models using the filtered PAV matrix. A regression line, with 95% confidence intervals was overlaid to view the general trend in gene presence with distance from Santiago.

The presence-absence binary matrix generated above was utilised to assess phylogenetic relationships with the scikit-bioPython package (v0.7.0) DistanceMatrix() function, run using a custom Python script [GitHub]. The Neighbourhood joining tree was generated using hamming’s distance, with the PAV matrix as input, to assess similarity of presence/absence variation profiles between individuals. [more info needed].

**Data Availability statement.**

All custom scripts mentioned in the above methods are stored on GitHub at this repository: [FINISH]

**Supplementary Data**

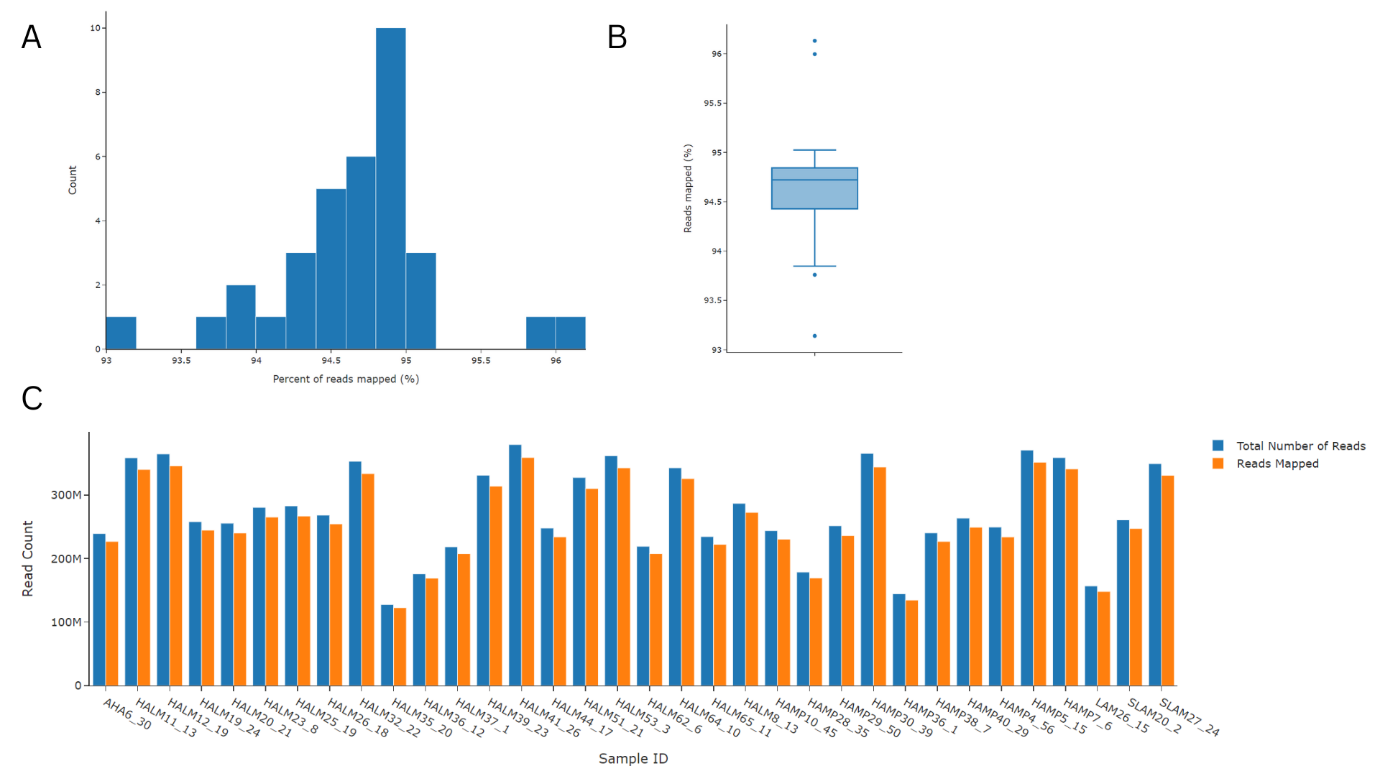
**Table 1.** The thirty-four *Scalesia* individuals used to build the pangenome and associated metadata.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No.** | **Individual ID** | **Island** | **Species** | **Latitude** | **Longitude** | **Habitat** |
| 1 | HALM35\_20 | Bartolomé | *S. stewartii* | -0.28411 | -90.5538 | Lava |
| 2 | HALM32\_22 | Santiago | *S. pedunculata 2* | -0.23161 | -90.7563 | Evergreen Forest and Shrubland |
| 3 | HALM36\_12 | Santiago | *S. stewartii* | -0.28942 | -90.5743 | Deciduous Grassland |
| 4 | HALM44\_17 | Isabela | *S. cordata* | -0.84669 | -91.0219 | Agricultural Area |
| 5 | HALM51\_21 | Isabela | *S. microcephala* | -0.19053 | -91.2959 | Lava |
| 6 | HALM53\_3 | Isabela | *S. cordata* | -0.89442 | -91.0021 | Evergreen Forest and Shrubland |
| 7 | AHA6\_30 | Pinzon | *S. baurii subsp. baurii* | -0.60639 | -90.67 | Deciduous Forest |
| 8 | HALM37\_1 | Isla Eden | *S. aspera* | -0.56008 | -90.5333 | Deciduous grassland |
| 9 | HALM26\_18 | Santa Cruz | *S. retroflexa* | -0.74464 | -90.2728 | Deciduous Forest |
| 10 | HALM39\_23 | Santa Cruz | *S. crockeri* | -0.57919 | -90.173 | Deciduous grassland |
| 11 | HALM41\_26 | Santa Cruz | *S. cfr. retroflexa* | -0.69469 | -90.1967 | Deciduous Forest |
| 12 | HALM62\_6 | Santa Cruz | *S. helleri* | -0.75431 | -90.4241 | Deciduous Forest |
| 13 | HALM64\_10 | Santa Cruz | *S. aspera* | -0.60156 | -90.5352 | Deciduous Forest |
| 14 | HALM8\_13 | Santa Cruz | *S. aspera x S. crockeri* | -0.53922 | -90.3171 | Deciduous Forest |
| 15 | HAMP10\_45 | Santa Cruz | *S. affinis* | -0.73944 | -90.2658 | Deciduous Forest |
| 16 | HAMP4\_56 | Santa Cruz | *S. affinis* | -0.73778 | -90.3244 | Deciduous Forest |
| 17 | HAMP5\_15 | Santa Cruz | *S. crockeri* | -0.48992 | -90.2803 | Deciduous Forest |
| 18 | HAMP7\_6 | Santa Cruz | *S. aspera x S. crockeri* | -0.54667 | -90.3192 | Deciduous Forest |
| 19 | LAM26\_15 | Santa Cruz | *S. retroflexa* | -0.71295 | -90.2269 | Deciduous Forest |
| 20 | HALM25\_19 | Santa Cruz | *S. pedunculata 2* | -0.62417 | -90.3839 | Evergreen Forest and Shrubland |
| 21 | HALM65\_11 | Santa Fé | *S. helleri* | -0.80444 | -90.0417 | Deciduous Forest |
| 22 | SLAM20\_2 | Pinta | *S. baurii subsp. hopkinsii* | 0.541572 | -90.7411 | Lava |
| 23 | SLAM27\_24 | Pinta | *S. baurii subsp. hopkinsii* | 0.573333 | -90.7644 | Evergreen Forest and Shrubland |
| 24 | HALM11\_13 | San Cristóbal | *S. divisa* | -0.79339 | -89.3355 | Deciduous Forest |
| 25 | HALM12\_19 | San Cristóbal | *S. divisa* | -0.77961 | -89.4114 | Deciduous Shrubland |
| 26 | HALM19\_24 | San Cristóbal | *S. divisa* | -0.71503 | -89.3888 | Lava |
| 27 | HALM20\_21 | San Cristóbal | *S. incisa* | -0.69567 | -89.3009 | Lava |
| 28 | HALM23\_8 | San Cristóbal | *S. divisa x S. incisa* | -0.81861 | -89.3533 | Deciduous Forest |
| 29 | HAMP36\_1 | San Cristóbal | *S. pedunculata 1* | -0.90111 | -89.4397 | Agricultural Area |
| 30 | HAMP38\_7 | San Cristóbal | *S. gordilloi* | -0.93028 | -89.6081 | Deciduous Forest |
| 31 | HAMP40\_29 | San Cristóbal | *S. incisa* | -0.71167 | -89.2475 | Deciduous Grassland |
| 32 | HAMP28\_35 | Floreana | *S. villosa* | -1.22528 | -90.4278 | Deciduous Shrubland |
| 33 | HAMP29\_50 | Floreana | *S. affinis* | -1.28278 | -90.4914 | Deciduous Forest |
| 34 | HAMP30\_39 | Floreana | *S. pedunculata 1* | -1.29194 | -90.4531 | Evergreen Forest and Shrubland |

**Table 2.** All significantly enriched go terms within the non-core gene set with associated uncorrected p value and FDR score.

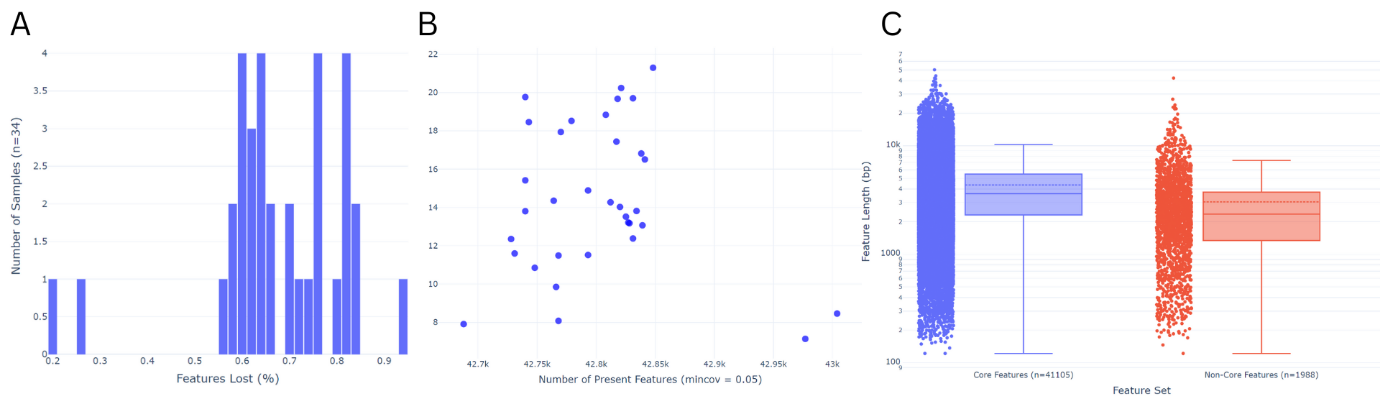


**Figure 1:**



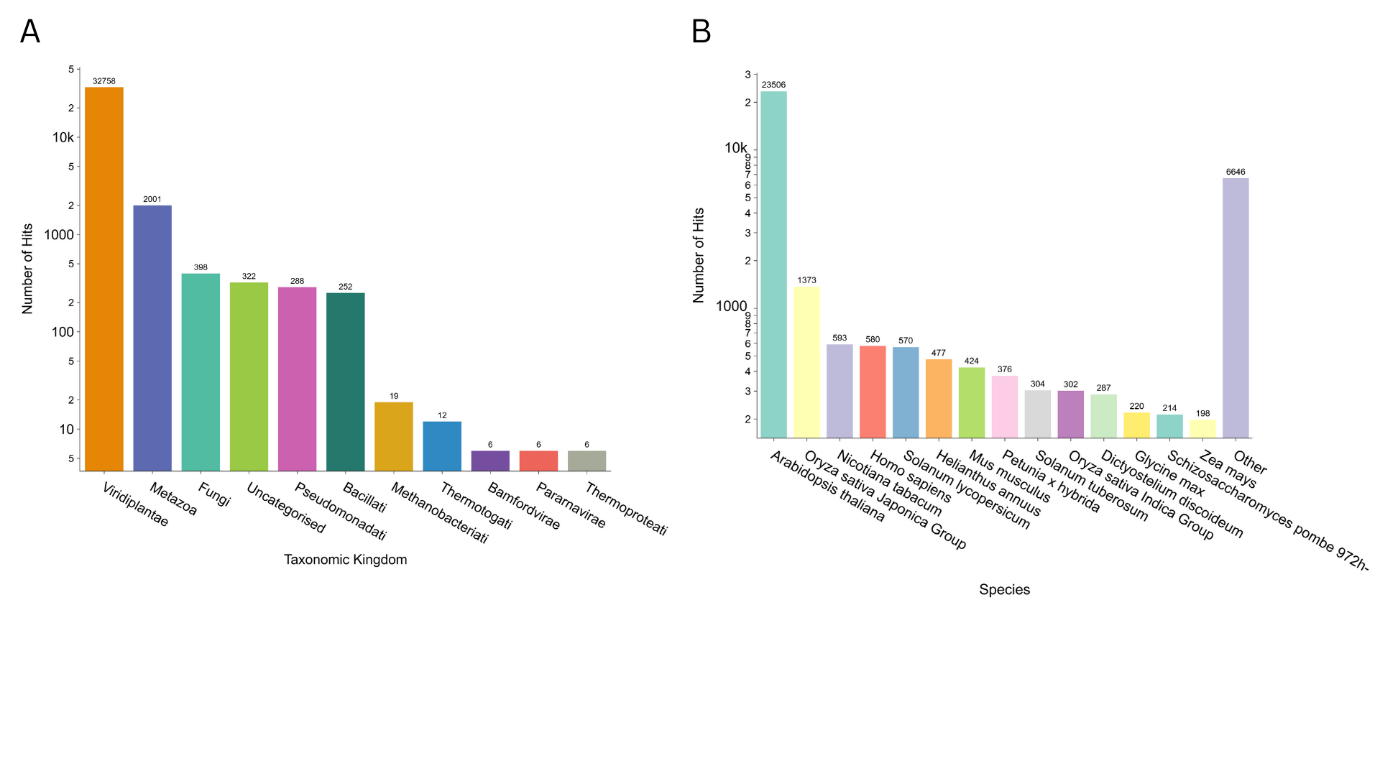
1. Histogram summarising the percentage of reads mapped to the reference genome (n=34), bin width = 0.5%.
2. Box plot representing the percentage of reads mapped to the reference genome. medium (n=34).
3. Bar plot showing total number of reads for each individual (blue bars) and the total number of reads mapped to the reference genome during pangenome assembly (orange).

**Figure 2:**



1. Histogram summarising the percentage of GFF file features marked as abcent via the sgsgeneloss package (mincov=0.05), n=34, bin width = 0.02%.
2. Scatter plot showing average read depth (Y) against number of features marked present for each individual (X) using the sgsgeneloss package (mincov 0.05).
3. Box plot + scattergram of core and variable feature datasets (variable = features absent in at least one individual). Each feature plotted on Y axis and associated length in base pairs (BP).

**Figure 3.**

****

(a) DIAMOND BlastP hits by taxonomic kingdom for functional annotation of 43,093 reference GFF features. Only the top hits are shown. Y axis log 10 scaled.

(b) DIAMOND BlastP hits by species for functional annotation of the reference GFF features. Y axis log 10 scaled. “Other” defines all species other than those with the highest fourteen counts binned as one bar.

**Figure 4.**

**A screen shot of a graph

AI-generated content may be incorrect.**

1. Scatter plot and regression line for all variable PAV’s as described in figure 4A.
2. Python Great table showing total number of present genes per individual, in 1) entire variable dataset. 2) subset of noncore dataset containing genes mapping to at least one of the top 10 most enriched GO terms with biological process namespace.

**References:**

OboNet: <https://zenodo.org/records/15086462>

GeoPandas: <https://doi.org/10.5281/zenodo.3946761>