Pangenome analysis highlights 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 all eighteen *Scalesia* species, collected among ten islands of the Galápagos. We ran presence/absence variation analysis to identify 1988 variable gene regions within the sample set, mapping these loci to 1117 unique genes. Within these variable regions, we identified a significant loss of genes involved in Germacrene A derived terpenoid biosynthesis, in addition to other secondary metabolic functions. This work represents the first study on gene loss within *Scalesia* and identifies gene loss as a contributing genetic mechanism in the formation of the “plant island syndrome” phenotype. We show differences in presence/absence profiles of individuals within the Galápagos varies between islands, and how this phenomenon can be used to support the current migratory theory of adaptive radiation in *Scalesia* across the archipelago*.* This work implies gene-absence profiles between sub-populations can be utilised 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, is 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* species 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 deemed *Scalesia* a valuable system 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 for gaining new 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 fifteen *Scalesia* species, estimating time passed since divergence of *Scalesia* from its South American relatives the subsequent diversification of the genus throughout the archipelago. They published evidence 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 the eighteen *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 been put forward, with the publication of the critically endangered *Scalesia atractyloides* genome (10). This work provided valuable advances into the genetic profile of the *Scalesia* genus common ancestor, by 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 key to adaption and diversification. These limitations are amplified in species produced through adaptive radiation, in which evolutionary novelty may arise through lineage-specific gene content.

Pan-genomics has emerged to overcome these limitations; it’s goal to capture all genomic content within a population of a taxonomic group. This includes core genetic material shared by all individuals, as well as accessory, non-core, material only present in a subset (11). The value of pan-genomics has been demonstrated in species such as *Brassica Oleracea*, in which extensive gene PAV linked to cultivar traits (12). In soybean and tomato, pangenomes uncovered novel genes absent from their relative reference genomes that were associated with disease resistance and environmental adaption (13) (14). These case-studies illustrate how population-level pan-genomes can expose functionally and evolutionary important genomic regions, overlooked by traditional, individual-based genomics. Pan-genomes enable the study of gene loss, that has long been recognised as a contributing factor for diversification of species (15). However, it’s role and importance in driving adaptive radiation of plant-island species remains understudied. Plant-island species often face strong ecological pressures, such as drought, nutrient limitations and isolation, which shape their genomic architecture (10) (16). This observation has led to the formulation of the “plant-island syndrome” hypothesis – in which such species undergo convergent evolution, manifesting in traits including increased size, woodiness and reduced defences against herbivores (17). These 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 loss has had in driving the formation of these phenotypes has not yet been documented.

We aimed to build a genus-level pangenome of all eighteen species of *Scalesia* utilising data obtained from [where], employing the iterative mapping approach (18). Using these data, we endeavoured to gain understanding of how genomic PAV may contribute to *Scalesia*’s adaptive radiation across the Galápagos. This study presents the first observations of gene loss within *Scalesia* and that the phenomenon has 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 well known to be involved in defence against herbivores (19,20). Additionally, we attempted to reveal whether relative differences in gene loss profiles of *Scalesia* individuals can be used to support the current migratory theory of the genus across the Galápagos as presented by Fernández-Mazuecos *et al* (2020) (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 also show that the extent of gene loss in *Scalesia* generally increases with distance from the central islands of the Galápagos. This pattern of gene loss supports the current theory of *Scalesia* migration by adaptive radiation over 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 critically endangered *Scalesia* species may help to inform conservation strategies and preserve biodiversity within the Galápagos.

**Materials and Methods:**

**Pangenome Assembly.**  
A pre-published genome for *Scalesia atractyloides* was used as the reference dataset for this study, obtained from the existing DOI: Dryad: 10.5061/dryad.8gtht76rh (10). The genome FASTA file was indexed with Bowtie2 (v2.4.5) (21) and used for reference-based alignment of each of the thirty-four *Scalasia* genomes as detailed below. The *Scalasia* pangenome was assembled using the iterative mapping approach (18). This protocol has been used to generate pangenomes for multiple plant species in the past (22,23). Illumina paired end reads were obtained from [complete] for thirty-four individuals within the Galápagos Islands, collectively representing eighteen different *Scalasia* species [Sup. table 1]. Initial read quality metrics were obtained using FastQC (v0.11.9). Paired samples were processed using Trimmomatic (v0.39) (24) with the following parameters:  
 *`*ILLUMINACLIP:$ADAPTERS:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:99 TOPHRED33*`*.

Post-trimmed quality was assessed with FastQC/MultiQC (v1.12) (25). Trimmed FASTA files were iteratively aligned to the reference genome using Bowtie2 (v2.4.5) (21) with the --sensitive flag and insert size parameters: -I 0 -X 1000. Sorted BAM files for the thirty-four samples were generated with Samtools (v0.5.0) (26) using *`*view*`*, *`*sort*`* and `index*`* commands (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 sequentially extracting unmapped reads from each BAM file using Samtools (-f 4) flag. Unmapped read headers were used to retrieve their relative uncleaned counterparts and assembled *de novo* with MaSuRCA (v4.14) (27). BBmap (v39.33) (28) was used to estimate average insert size/standard deviation of samples. (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. Hits with an evalue of < 1e-10 were removed from further processing. The filtered assembly was mapped back to the original *Scalasia* reference using Minimap2 (v2.30) with default parameters (30) .Unmapped contigs (57783) were extracted with Samtools (f -4) and queried against the core-nt database [more info]. Taxonomy information was mapped to blastn hits with Taxonkit (v0.20.0) (31). Contigs with hits belong to “streptophyta” – the green plants kingdom and “novel” contigs returning no hits were retained. This file, containing 6338 contigs/2352 603 base pairs was concatenated to the reference genome to build the final pangenome. Hits per species and kingdom were visualised with Python Plotly [Sup figure. 3].

**[ 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 the Sgsgeneloss (v0.1) (32) package to calculate presence/absence variation of each sample against the reference dataset on the thirty-four *Scalesia* BAM files. Sgsgeneloss was run using a minimum coverage (mincov) value 0.05, this value has become standardised across other gene-loss studies (33,34). We generated presence/absence calls of the thirty-four samples 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 sample contained adequate read depth/coverage to prevent the calling of false absence of features (Sup. Fig2). All thirty-four samples 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 sample, before merging with Pandas (v2.2.3) (35) to generate the binary PAV binary matrix. Non-core 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 depth of coverage was sufficiently high (~10x coverage 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 non-core (missing in at least one individual). The non-core dataset accounted for 4.6% of the 43093 total features within the reference GFF file.

**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 sample metadata (Sup. Table 1).

**Functional Annotation of The *Scalesia atractyloides* Reference GFF.**  
Protein sequences were extracted for each feature of the *Scalesia atractyloides* GFF 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, using the following code snippet to return all hits, paired with associated taxonomy information:  
  
*`*diamond blastp -q protein\_clean.fa -d uniprot\_sprot.dmnd -o diamond\_results.tsv -p 64 --sensitive -e 1e-5 -k 1 --outfmt 6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore staxids stitle sscinames sskingdoms skingdoms sphylums*`*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. We were able to obtain high-confidence hits for 36070/43093 of the total GFF file features. For the non-core feature set, we obtained high-confidence hits for 1508/1988 features. This subset contained a total of 1117 unique genes.

**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 sample 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 reference point was chosen as the geographic centre for all distance-based analyses.

We assessed association of presence/absence and geographic distance from Santiago within the non-core gene set by running Ordinary Least Squares (OLS) and Poisson regression analysis using the statsmodels (v 0.14.5) (43) Python package. Both models did not identify a significant relationship between distance from Santiago and number of absent genes [SF4.1]. The non-core gene PAV matrix was filtered to identify genes that mapped to the top 10 GO terms flagged as overabundant (absent) in the GO enrichment study based on P-value. Total number of present genes for this subset were calculated for each sample 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, both models identified a significant negative relationship between distance from Santiago and gene absence for the top ten go IDs – with P < 0.001 and P = 0.037 respectively. A regression line, with 95% confidence intervals was overlayed to view the general trend in gene presence with distance from Santiago Island.

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]. To generate a second tree... [mitch to finish – mashplot]

**Custom Scripts.**

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

**Results**

**General MAKER Pangenome results.**

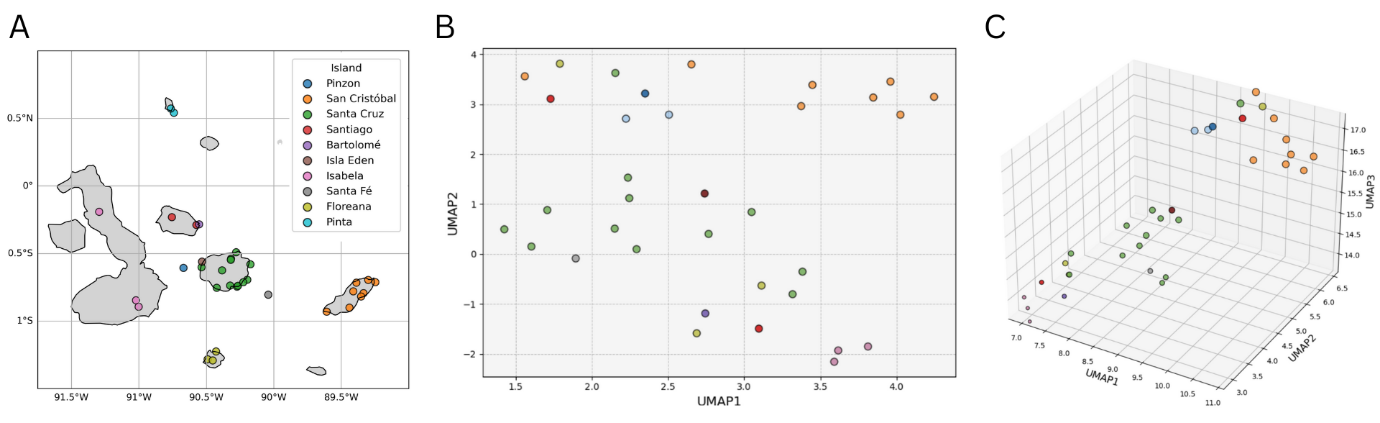
[To complete]

**UMAP Analysis Indicates Spatially Close Individuals Share Similar Feature Absence Profiles.**

To investigate presence/absence variation of the *Scalesia* Genus within the Galápagos, we assembled genome sequences for thirty-four *Scalesia* individuals collected from 10 different islands around the Galápagos Archipelago [F1.1]. These samples collectively represent a dataset of eighteen different *Scalesia* species. PAV analysis revealed 1988/43093 features defined within the reference GFF as non-core (missing in at least one individual).

To assess the extent and pattern of feature absence variation between individuals, we employed UMAP clustering to highlight differences in absence variation between individuals. We observed individuals tended to cluster with other samples from the same island or island group. For example, all three samples from island Isabela clustered together, away from samples obtained from San Cristóbal – located at the Eastern end of the archipelago. UMAP clustering also revealed many samples from island Santa Cruz formed a distinct cluster, separate from most samples of San Cristóbal, and samples from the Westerly islands of Santiago and Isabela. Despite forming rough clusters, several samples did not fit the overall pattern. For example, one sample from Santiago had a PAV profile more representative of samples from San Cristóbal, rather than other samples taken from the same or 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 clustering with samples from the central islands – Isabela, Santiago/ Bartolomé and Santa Fe.

**Figure 1.**



1. Map of the Galápagos Islands with longitude/latitude coordinates of samples obtained for the study. Colour indicates which island the sample 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; however, their relative positions can be visualised using sample co-ordinates..
2. 2D 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.
3. 3D UMAP of PAV matrix, using identical parameters to the 2D UMAP (b), point size scaled using UMAP-Z co-ordinate to enhance depth perception.

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

Functional annotation of the GFF file matched 36550/43093 features to genes (E-value < 1e-5) using BLAST. Go enrichment analysis of the annotated non-core gene set (n=1508) identified 45 significantly enriched GO terms (P <0.01). The top 10 most enriched GO terms for each GO namespace are shown in Figure 2B. The heatmap indicates PAV of each gene within the sample set, sorted by co-ordinate. Results indicated a general loss of genes involved in secondary metabolite biosynthesis. More specifically, the study highlighted multiple terms 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 non-core subset implies *Scalasia* species may be downregulating production of these compounds via gene loss. Figure 3 shows all genes within the non-core dataset that map to go term GO:0016114 - terpenoid biosynthesis. This term was the most significantly enriched biological process identified in the study. This table contains of fourteen unique genes across unique 35 loci within the genome.

**Figure 2.**

A green and blue graph

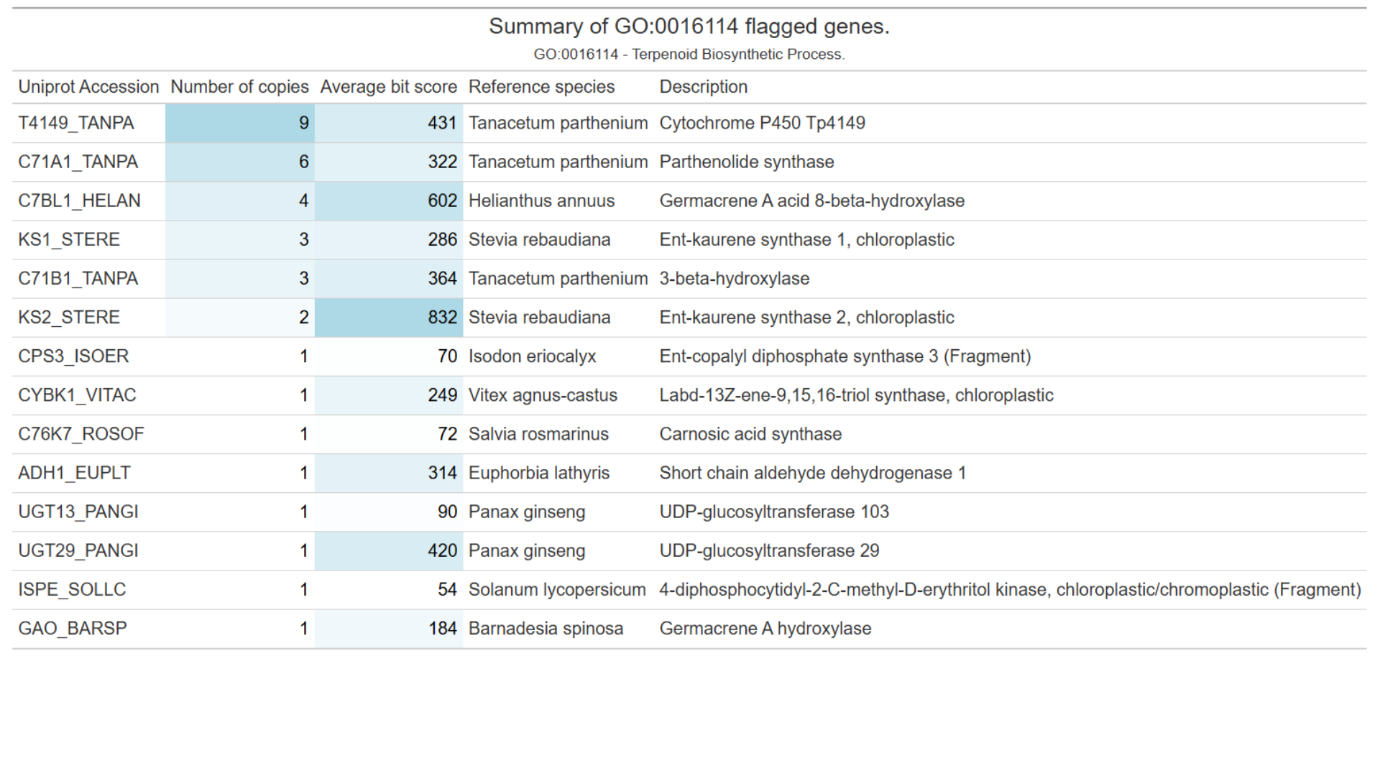
AI-generated content may be incorrect.

1. Heatmap representing presence (green) / absence (white) variation of genes mapping to at least one of the top five most-significantly enriched GO biological processes as shown in (b). Y axis sorted by start position of loci within the linearised reference genome, these co-ordinates were obtained using the reference GFF file.
2. Results of the GO enrichment analysis on the non-core gene set (genes missing in at least one individual). Results are separated by GO namespace. Go terms were mapped to human readable descriptions.

Amongst the genes identified was Parthenolide synthase (C71A1\_TANPA), returning high-confidence hits to 6 unique loci within the non-core subset. This enzyme moderates the conversion of costunolide to parthenolide, responsible for the final step in parthenolide biosynthesis from a Farnesyl pyrophosphate (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 2A indicates many samples are missing C71A1\_TANPA at one locus, with a subset of samples primarily obtained from San Cristobal (orange), showing similar absence of this gene at 3 loci clustered together, except for HALM37\_1 (Isla Eden), all other samples were flagged as present for these loci. C7BL1\_HELAN (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 (48). This enzyme returned high-confidence hits to four different loci within the non-core gene set. Visualisation of the presence/absence pattern of these loci indicated varying degrees of absence between samples. One locus indicated heavy absence of the gene across samples from San Cristobal, whilst absence was less common among the central islands. Samples from Isabela, Santiago and Bartolomé all retained presence at this locus (except HALM32\_22, from Santiago).

T4149\_TANPA presented high-confidence hits to nine different loci within the non-core gene set. It has been suggested that this enzyme also plays a role in the same biochemical pathway, however it’s exact function remains undetermined (45). Figure 2A indicated that these loci clustered at 2 regions within the genome, at one of these regions, including 5/9 gene copies, absence of these features only occurred in two samples from Isabela, with remaining samples showing uniform presence. We also identified gene absence of two Ent-kaurene synthase’s – KS1\_STERE and KS2\_STERE, at 3 and 2 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 biochemical pathway (50).

**Figure 3.**



**Figure 3**. Great table (ref) representing all unique genes within non-core gene set mapping to GO term 0016114 – Terpenoid biosynthesis. Number of copies represents the number of loci within the non-core feature set with high confident hits to the relative accession. Average bit score represents the mean bit-score as obtained from the `blastp`- 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.]

[Could add row for number of core loci in table too?]

**Spatial Variation in Gene Presence–Absence Reflects Progressive Gene Loss Away from Galápagos Central Islands.**  
We calculated distance in longitude/latitude of each sample 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 non-core 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 and Poisson), P < 0.001 and P = 0.037 respectively.

**Figure 4**.

A diagram of a graph

AI-generated content may be incorrect.

1. 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. Samples are coloured by island using key in Figure 1.A. A Dashed best-fit linear regression line across all points with 95% confidence intervals was overlayed to show trend.
2. Neighbour-joining tree calculated using Hamming’s distance on the Non-core PAV matrix. Samples coloured according to key in Figure 1.A. [boostrap for support?]

[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?

[Where to put neighbourhood joining trees]

[Mainland species – was too diverged to accurately call presence/absence using the reference dataset]. – Maybe another option would be to call PAV using it as the reference? Would require a larger chunk of work.

**Discussion.**

**Maker Results Paragraph:**

* Does the number of non-core genes reflect that of other plants?
* High number of gene repeats implications.
* High level of contamination in samples.

The pangenome constructed from thirty-four *Scalesia* individuals identified 1988 previously uncharacterised variable regions within the genus. Analysis of these regions identified 1117 unique genes that may contribute to the diversification of different *Scalesia* species. Initial UMAP analysis indicated a rough clustering of samples according to island. However, some samples exhibited PAV profiles more characteristic of those in geographically distant islands. For example, HALM32\_22 (S. *pedunculata*) from Santiago, clustered with samples of San Cristóbal, rather than with geographically nearer samples of Santiago and Bartolomé. This pattern is maintained in the neighbour joining tree (Fig 4.B), also showing HALM32\_22 groups with individuals from San Cristóbal, rather than Santiago or neighbouring islands. In addition, individual HALM25\_19 also revealed a comparatively different PAV profile compared with other samples obtained from Santa Cruz, clustering close to HALM32\_22 (S. *pedunculata*) 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 residing 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 data is required to confirm this hypothesis.

Go enrichment analysis of functionally annotated non-core 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) . When taken in context with other observations, including the reduction in the number of large herbivores within the Galapagos, compared with the mainland (52). It is evolutionary sensible for *Scalesia* to downregulate production of herbivore defensive compounds. The observation that soil quality and mineral availability varies considerably within the Galapagos also supports this finding (53). This phenomenon may have contributed to *Scalesia* streamlining metabolism through gene loss, downregulating expression of metabolically expensive secondary compounds such as terpenoids. Interestingly, Strahlhofer *et al* (2021) observed the older island San Cristóbal had noticeably poorer soil fertility than younger islands of the archipelago, whilst we highlighted increased gene loss on San Cristobal in this study.

Gene loss relating to Germacrene A terpenoid biosynthesis strongly suggests a downregulation of terpenoid synthesis, due to loss of multiple genes involved in the same biochemical pathway - Parthenolide synthase, Germacrene A acid 8-beta-hydroxylase and potentially T4149\_TANPA. However, it would be premature to assume Germacrene A based terpenoid biosynthesis is completely lost in *Scalesia,* as the repetitive nature of these genes implies not all expression is lost. When we confirmed how many loci mapped to each of these genes using the reference GFF file, we found some copies of these genes mapped to core regions, implying gene loss has not occurred at these loci [Add to figure 3?]. This observation implies it is more probable of diversity in the levels of sesquiterpene lactone production across *Scalasia* species of the Galápagos, rather than universal loss of the biological process altogether. The observation of gene loss in KS1\_STERE and KS2\_STERE enzymes, involved in gibberellin biosynthesis was surprising, as Darwin’s Daises 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 study is required to identify why KS1\_STERE and KS2\_STERE genes are undergoing gene loss. These data may hint at secondary, undocumented functions for these enzymes, in processing intermediaries of Germacrene A derived terpenoids. Reduction in biosynthesis of defensive compounds is a classic trait of “plant-island syndrome” (17). These findings suggest that in *Scalesia*, gene loss is a likely underlying genetic mechanism responsible for the formation of the phenotype – via downregulation of Germacrene A derived terpenoids.

In line with previous evidence, we hypothesised the pattern of feature absence observed may be associated with the evolutionary migration pattern of *Scalasia* across the Galápagos. Using prior work by Fernández-Mazuecos *et al (2020)*, suggesting a common ancestor of *Scalasia* 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 entire non-core gene set, we did not identify a significant trend [Sup Fig 4.A]. Sub-setting the non-core PAV matrix to retain only genes mapped to at least one of the top 10 most enriched GO:BP terms revealed significant a negative correlation in multiple models [Fig 3.A]. Exactly why this subset caused exaggeration of the trend and it’s subsequent significance remains to be determined. An explanation for this observation may be that GO term filtering reduced the number of false absences within the dataset, as enriched pathways are less likely to arise by chance, rather than individual false absence calls within the matrix. 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 samples of San Cristóbal, the number of present genes varied between 1220 and 1310 min/max respectively, resulting in a range of 90 PAVs. Even between samples of the same species [example]. 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 gene loss findings].

These results suggest gene loss within *Scalesia* has likely contributed to the formation of the “plant-island-syndrome” phenotype. We identified 1988 previously unidentified non-core regions of the *Scalesia* pan-genome and subsequent analysis of these regions uncovered evidence of loss of multiple genes involved in Germacrene A derived biosynthesis. Using these non-core regions, we highlighted PAV profiles of *Scalesia* individuals tend to vary less among geographically close samples 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 sample size of this dataset. In summary, this study demonstrates pan-genome 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.

**Supplementary Data**

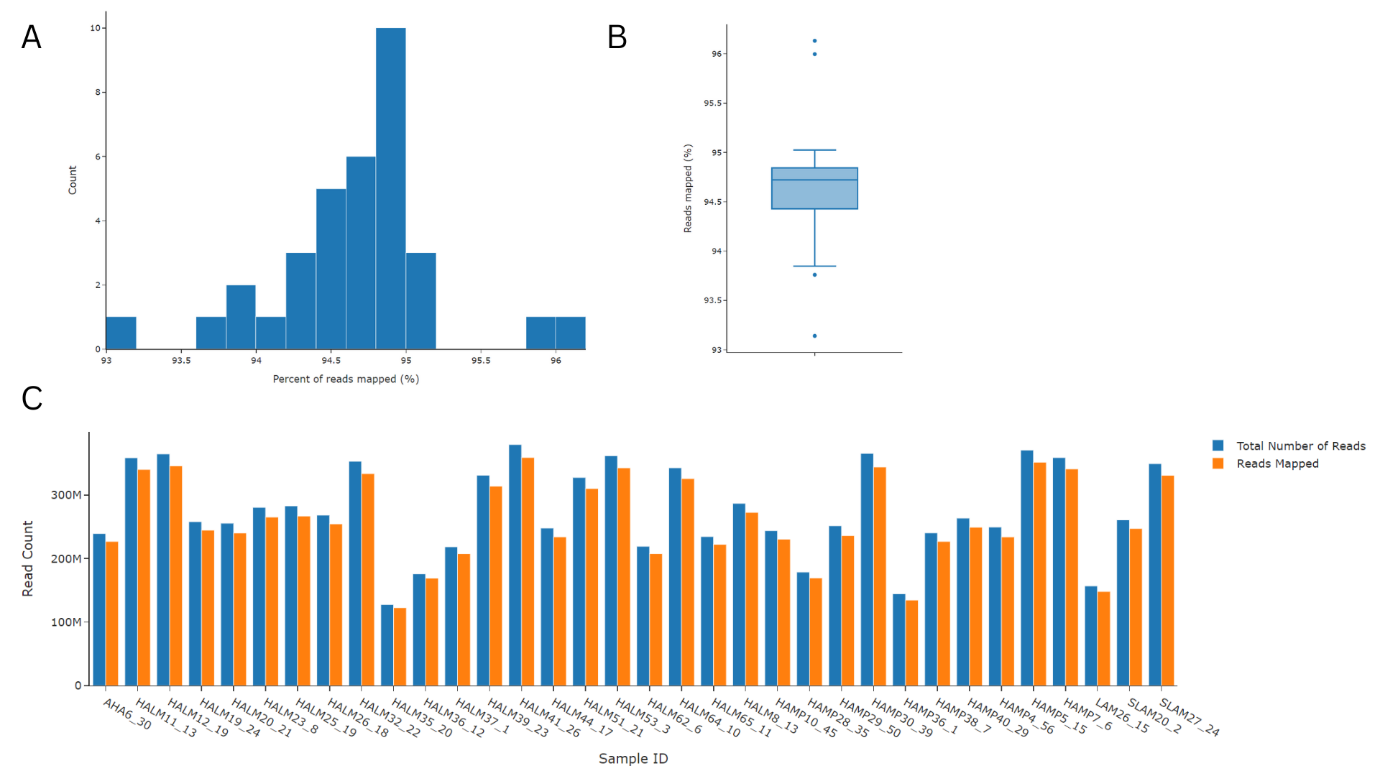
[Maybe add Teng’s UMAPs to show that we could not find clustering for any other metadata?]

**Table 1:**

The thirty-four *Scalesia* samples used to build the pangenome and associated metadata.

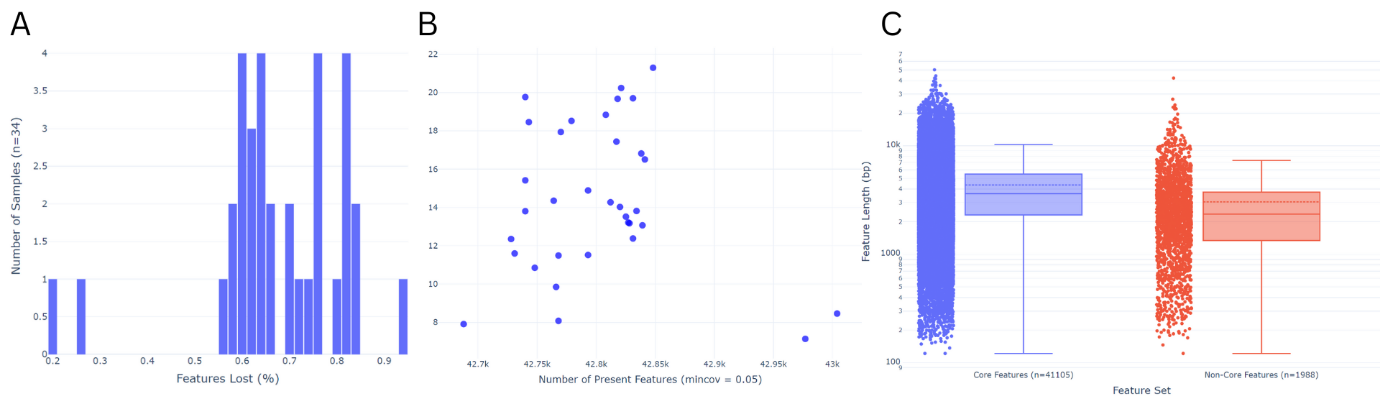
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample ID | Island | Species | Latitude | Longitude | Habitat |
| HALM35\_20 | Bartolomé | Scalesia stewartii | -0.28411 | -90.5538 | lava |
| HALM32\_22 | Santiago | Scalesia pedunculata 2 | -0.23161 | -90.7563 | Evergreen Forest and Shrubland |
| HALM36\_12 | Santiago | Scalesia stewartii | -0.28942 | -90.5743 | deciduous grassland |
| HALM44\_17 | Isabela | Scalesia cordata | -0.84669 | -91.0219 | agricultural area |
| HALM51\_21 | Isabela | Scalesia microcephala | -0.19053 | -91.2959 | lava |
| HALM53\_3 | Isabela | Scalesia cordata | -0.89442 | -91.0021 | Evergreen Forest and Shrubland |
| AHA6\_30 | Pinzon | Scalesia baurii ssp. baurii | -0.60639 | -90.67 | deciduous forest |
| HALM37\_1 | Isla Eden | Scalesia aspera | -0.56008 | -90.5333 | deciduous grassland |
| HALM26\_18 | Santa Cruz | Scalesia retroflexa | -0.74464 | -90.2728 | deciduous forest |
| HALM39\_23 | Santa Cruz | Scalesia crockeri | -0.57919 | -90.173 | deciduous grassland |
| HALM41\_26 | Santa Cruz | Scalesia cfr. retroflexa | -0.69469 | -90.1967 | deciduous forest |
| HALM62\_6 | Santa Cruz | Scalesia helleri | -0.75431 | -90.4241 | deciduous forest |
| HALM64\_10 | Santa Cruz | Scalesia aspera | -0.60156 | -90.5352 | deciduous forest |
| HALM8\_13 | Santa Cruz | Scalesia aspera x S. crockeri | -0.53922 | -90.3171 | deciduous forest |
| HAMP10\_45 | Santa Cruz | Scalesia affinis | -0.73944 | -90.2658 | deciduous forest |
| HAMP4\_56 | Santa Cruz | Scalesia affinis | -0.73778 | -90.3244 | deciduous forest |
| HAMP5\_15 | Santa Cruz | Scalesia crockeri | -0.48992 | -90.2803 | deciduous forest |
| HAMP7\_6 | Santa Cruz | Scalesia aspera x S. crockeri | -0.54667 | -90.3192 | deciduous forest |
| LAM26\_15 | Santa Cruz | Scalesia retroflexa | -0.71295 | -90.2269 | deciduous forest |
| HALM25\_19 | Santa Cruz | Scalesia pedunculata 2 | -0.62417 | -90.3839 | Evergreen Forest and Shrubland |
| HALM65\_11 | Santa Fé | Scalesia helleri | -0.80444 | -90.0417 | deciduous forest |
| SLAM20\_2 | Pinta | Scalesia baurii ssp. hopkinsii | 0.541572 | -90.7411 | lava |
| SLAM27\_24 | Pinta | Scalesia baurii ssp. hopkinsii | 0.573333 | -90.7644 | Evergreen Forest and Shrubland |
| HALM11\_13 | San Cristóbal | Scalesia divisa | -0.79339 | -89.3355 | deciduous forest |
| HALM12\_19 | San Cristóbal | Scalesia divisa | -0.77961 | -89.4114 | Deciduous Shrubland |
| HALM19\_24 | San Cristóbal | Scalesia divisa | -0.71503 | -89.3888 | lava |
| HALM20\_21 | San Cristóbal | Scalesia incisa | -0.69567 | -89.3009 | lava |
| HALM23\_8 | San Cristóbal | Scalesia divisa x S. incisa | -0.81861 | -89.3533 | deciduous forest |
| HAMP36\_1 | San Cristóbal | Scalesia pedunculata 1 | -0.90111 | -89.4397 | agricultural area |
| HAMP38\_7 | San Cristóbal | Scalesia gordilloi | -0.93028 | -89.6081 | deciduous forest |
| HAMP40\_29 | San Cristóbal | Scalesia incisa | -0.71167 | -89.2475 | deciduous grassland |
| HAMP28\_35 | Floreana | Scalesia villosa | -1.22528 | -90.4278 | Deciduous Shrubland |
| HAMP29\_50 | Floreana | Scalesia affinis | -1.28278 | -90.4914 | deciduous forest |
| HAMP30\_39 | Floreana | Scalesia pedunculata 1 | -1.29194 | -90.4531 | Evergreen Forest and Shrubland |

**Figure 1:**



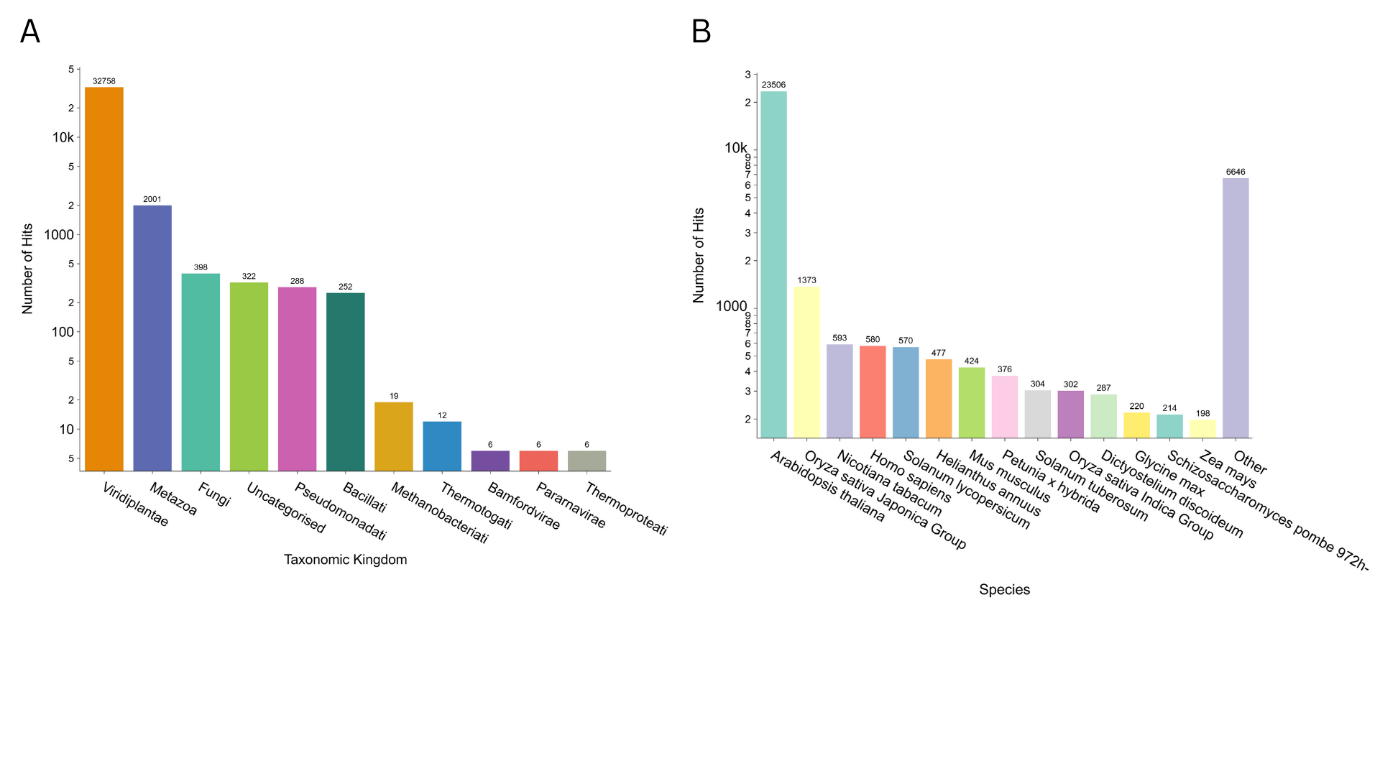
1. Histogram summarising the percentage of reads mapped to the reference genome (n=34), bin width = 0.5%.
2. Box plot representing the percentage of reads mapped to the reference genome. medium (n=34).
3. Bar plot showing total number of reads for each sample (blue bars) and the total number of reads 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 sample (X) using the sgsgeneloss package (mincov 0.05).
3. Box plot + scattergram of core and non-core feature datasets (non-core = features absent in at least one sample). 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 non-core PAV’s as described in figure 4A.
2. Python Great table showing total number of present genes per sample, in 1) entire non-core 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>