**Applying selection scans to identify candidate-genes implicated in enabling parallel adaptation of coastal environments by *Brassica fruticulosa***

Name: Matthew Gaskins

Names of other project members: Bailey Francis, Ana da Silva

**Abbreviations:**

AFD: Allele Frequency Difference

AGPs: Arabinogalactan-proteins

BLAST: Basic Local Alignment Search Tool

DDR: Diversity/Differentiation Residual

FST: Fixation Index

GO: Gene Ontology

LD: Linkage Disequilibrium

MAP: Microtubule-Associated Protein

NGS: Next-Generation Sequencing

NTP: Nucleoside Triphosphate

PCA: Principal Component Analysis

PGM: Population Genomic Metric

PI3K: Phosphoinositide 3-Kinase

ROS: Reactive Oxygen Species

SOS: Salt Overly Sensitive

SNP: Single-Nucleotide Polymorphism

**Project Role:**

Alongside Ana, I modified the demography script provided by Dr. Yant to generate PCAs and a neighbour-joining tree. I then employed VCFTools to generate pairwise-FST values between three geographically proximate population-pairs, and developed a custom R-script to identify the upper 99th-percentile of FST values from each population. I annotated FST outliers from both my VCFTools approach and Bailey’s ScanTools pipeline using SAMTools and BLAST. I also created R-scripts to both identify outliers using the DDR PGM, and to generate AFD plots within relevant candidate-genes.

**Aims and Experimental Question:**

The ability of plants to tolerate high environmental salinity facilitates their colonisation of physiologically stressful locations. Improving understanding of mechanisms by which halophilic plants confer salt-stress tolerance may rationalise crop design that improves yield from agricultural regions with high sodium concentrations. This report aims to elucidate the genetic basis underlying halotolerance within three populations of *Brassica fruticulosa* that colonise coastal environments. The initial objective is to infer demography between these three coastal populations and their neighbouring inland strains - characterising interpopulation relationships will determine whether strains are compared in three geographically-proximate pairs or two independent coastal/inland groups. Our report then employs two approaches to identifying frequently-observed genomic windows containing significant SNP-variation between closely-related coastal and inland strains - these genomic regions are implicated in functionally consequential adaptation facilitating halotolerance.

Table

Description automatically generatedWe aim to achieve our objective of interpreting intraspecies demography through PCA of putatively neutral fourfold-degenerate sites from our six populations. Table 1 provides locations and soil sodium-concentrations for each sample from investigated populations. This report further interrogates gene-flow between our six populations using SplitsTree to ultimately determine the nature of pairwise comparison in subsequent selection-scans.

Once demographic relationships between populations are characterised, we aim to identify genomic variation conferring halotolerance through pairwise-FST values calculated using VCFTools[1]. Here, we analyse a 65Mb-subset of the pre-existing 612MB-*B.fruticulosa* assembly constructed from Oxford-Nanopore long-read data using Redbean[2].We aim to distinguish the top 1% of FST-outliers between coastal and inland populations using an R-script identifying 5-kilobase windows containing minimum 25-SNPs. These residuals are compared to equivalent outlier windows identified using a second approach - ScanTools (adapted from <https://github.com/mbohutinska/ScanTools_ProtEvol>), to determine genomic-variation commonalities most likely to allow *B.fruticulosa* to withstand high sodium-concentrations. Functional characterisation of candidate-genes allows successful achievement of the primary aim of this report: enhanced understanding of genomic mechanisms conferring halotolerance in plants.

**Background and Motivation:**

Chart, waterfall chart

Description automatically generatedAmidst ever-decreasing arable land availability, agriculturally significant *Brassicae* represent established targets for design of crops more tolerant to poor-quality soil. One such species is *Brassica fruticulosa*, a wild diploid-outcrosser located primarily across inland regions of Southern-Europe, with scattered incidence along coastlines including the Mediterranean [3]. *B.fruticulosa* colonises at least nine distinct Spanish coastal locations - this consistent adaptation to high-saline environments represents a suitable paradigm for elucidating genomic mechanisms underlying halotolerance. Figure 1 displays increased ability of coastal *B.fruticulosa* to withstand high sodium-concentrations compared to inland relatives, with growth-reduction less significant in saline than non-saline strains.

In this study, we examine three geographically-proximate inland-coastal pairs - BRUC/GAR, PAU/LLAN and SOL/SFG2, listed in order of increasing salt-contrast. High-salinity represents environmental stress for plants, with particularly destructive effects on osmotic-regulation, therefore coastal populations must evolve mechanisms to tolerate this condition[4]. Frequent adaptive processes include enhanced sodium-efflux through SOS-pathway upregulation, elevated nitric-oxide production to mitigate disrupted photosynthetic ability, and synthesis of hormones regulating biotic-stress response[4]. It is reported *Brassicae* tolerate salt through osmolyte-accumulation and efficient sodium extrusion, however no functional evidence exists for *B.fruticulosa,* instigating this study[5].

To direct the nature of interpopulation comparison, PCAs reduce total SNP-variance into fewer dimensions that readily display interrelation between strains. Recent evolutionary divergence is established through proximal clustering of populations along significant principal-components. SplitsTree also allows the extent of gene-flow between populations to be inferred[6]. Analysing genetically similar populations mitigates noise from evolutionary-adaptation independent of halotolerance mechanisms.

Once relationships are identified between our six populations, pairwise-comparison of SNP- frequency is conducted using PGMs FST and AFD. Elevated footprints of differentiation between multiple coastal-inland pairs distinguish genomic regions important to phenotypic contrast - increased allele frequency in adapted populations indicates selective-sweep. However, one completely-differentiated SNP may simply reflect genetic drift. FST values are therefore averaged across longer windows, with varied genes distinguished through hitchhiking - recombination is unlikely to separate linked-genes given recent ecotype divergence[7]. Functional characterisation of genes under parallel-selection is important for elucidating halotolerance mechanisms in coastal populations. GO-analysis of predicted candidates will identify altered biological-processes and molecular-functions.

**Methods:**

A prefiltered subset of 1,909,722 biallelic-SNPs from five individuals of our six *B.fruticulosa* populations, three inland (BRUC,PAU and SOL) and three coastal (GAR,LLAN and SFG2), was provided by Levi Yant. We employed multiple complementary approaches using 159,720 putatively neutral fourfold-degenerate SNPs extracted from this dataset to infer genetic distances between populations. Using custom R-scripts, we employed adegenet(version2.1.1) to conduct both PCA and construct a neighbour-joining dendrogram based on estimated Nei genetic-distances between populations (with 1000 bootstrap-replicates)[8]. We then applied SplitsTree(version4.17.1) to visualise gene-flow between populations[6]. Mean Tajima’s-D values were calculated in 5-kilobase windows using VCFTools(version0.1.16)[1].

We then extracted a 65Mb-subset of the pre-existing 712Mb Oxford-Nanopore long-read assembly constructed using Redbean (N50=121kb, 90%-complete BUSCOs) containing the thirty-nine largest contigs (all >1MB). Pairwise-FST values were calculated across non-overlapping 5-kilobase windows of this subsetted genome using VCFTools. The upper 99th-percentile of windows containing at least 25-SNPs from empirical FST distributions for each population-pair were then identified using a custom R-script. Spearman’s rank correlation-coefficient between FST and SNP-frequency was calculated using R-package ‘dplyr’. VCFTools was also used to calculate linkage-disequilibrium values dictating window-size in subsequent selection-scans.

Outliers were then mapped to a complete *B.fruticulosa* annotation (containing approximately 11,000-genes) to identify their corresponding gene (Appendix 1). Nucleotide sequence from this gene was then extracted from the reference *B.fruticulosa* genome with SAMtools(version1.9), and input to the *A.thaliana* database from Nucleotide BLAST, run under default parameters with blastn(version2.13.0)[9,10]. Homologous genes with >75% identity to *A.thaliana* were characterised as functionally equivalent, and their TAIR-IDs recorded in lists of candidate-genes for parallel differentiation.

We conducted GO-analysis of all candidates identified from any population, using PANTHER(version16) to identify enrichment in particular GO-annotations[11]. Only terms displaying 1.5x-fold enrichment were retained for analysis, provided they also contained minimum three outliers and p-values beneath 0.05 and 0.01 for molecular-function and biological-process categories respectively. We subsequently conducted AFD-plots for relevant selection-sweep candidates using a custom R-script and the orientation-file provided in Appendix 2. Interrogated candidates represent the three largest pairwise-FST values for GO-terms displaying significant enrichment.

We refined population-pair-specific gene-lists by only retaining outliers present in at least two pairs. However, as only three candidates were conserved between at least two pairs, we employed an alternative method for refining directional-selection candidates - the ScanTools pipeline, run under default parameters with identical window-sizes and SNP-numbers to our previous approach. A Venn diagram of outliers identified using both or either program was generated using InteractiVenn software[12]. Subsequent GO and AFD-analysis (as described above) for candidates shared between approaches implicated genomic regions in parallel differentiation.

**Code:**

All scripts can be found inside my GitHub repository [[https://github.com/stymwg/Project3](https://github.com/stymwg/Project31)).

**Results and Discussion:**

Map

Description automatically generated with low confidenceTo achieve our initial aim of interpreting sample demography, we conducted principal-component analysis indicating overall grouping of our six *B.fruticulosa* populations into three coastal/inland pairs, clustered dependent on location rather than tolerance to salt. Figure 2 demonstrates the location of spatially proximate saline/non-saline pairs across Catalonia. PCAs, represented in Figures 3a-e, were produced using 159,720 putatively neutral fourfold-degenerate SNPs from subset data. Silent-substitutions have little functional impact (ignoring codon-bias) and thus do not shape adaptation - their relative variation reflects the extent of divergence between populations, allowing underlying phylogenetic relationships to be uncovered. PAU5 was removed from all demographic analysis and selection-scans, as this sample failed NGS-pipe read-depth filtering.

Principal-components 1 and 2, which together explain 24.39% of total-variance, split populations into three groups dependent on location. Indeed, Figure 3a displays one PAU-sample clustering more closely with coastal-relative LLAN than its own population, demonstrating the extent of similarity between pair-members. Surprisingly, the axis separating SOL and SFG2 explains the smallest proportion of total-variance of any dividing locational pairs (PC5=4.09%). SOL and SFG2 possess strongest sodium-concentration contrast, therefore their indicated similarity implies variance is largely unexplained by high-saline adaptation. Distinct genetic-similarity between each population-pair suggests halotolerance represents the outcome of multiple independent adaptation-events within each pair.

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, radar chart, bubble chart

Description automatically generatedChart

Description automatically generated with low confidence

**F**

Table

Description automatically generatedMoreover, similar mean-FST and nucleotide-diversity values in Table 2 demonstrate comparable divergence quantities within all population-pairs, and suggest independent adaptation events in each population occurred at approximately equivalent time-points. Minimal nucleotide-diversity within coastal/inland pairs suggests coastal colonisation represents a recent, postglacial event, although may simply reflect low population sizes. Similar nucleotide-diversity and Tajima’s-D values within pairs provide no evidence of genetic-bottleneck associated with coastal inhabitation.

Diagram

Description automatically generatedSimilar patterns of spatial separation were observed through further phylogenetic analysis of the same fourfold-degenerate dataset. The neighbour-joining tree displayed in Figure 4 confirms close phylogenetic relationships between geographically-proximal strains with strong bootstrap-support. Figure 4 also indicates similar divergence times between coastal-inland population-pairs, with all phylogenetic bifurcations between pairs inside a range of 0.01 Nei-distance.

Diagram

Description automatically generatedFigure 5 suggests more gene-flow within PAU/LLAN than either BRUC/GAR or SOL/SFG2, which combined with low nucleotide-diversity and Tajima’s D-values may explain extreme spatial-clustering between PAU and LLAN in PCA. SOL/SFG2 also experiences intrapair gene-flow, coherent with intermediate nucleotide-diversity - and even displays moderate interpopulation exchange with PAU/LLAN. This interrelationship appears logical as SOL/SFG2 and PAU/LLAN strains represent the most spatially-proximate unpaired populations, both geographically and between PC1-PC2. Despite BRUC/GAR diverging most recently, this pair contains largest nucleotide-diversity and pairwise FST values. BRUC and GAR represent the most geographically separate intrapair members, therefore correspondingly decreased gene-flow allows greater phylogenetic disparity despite a shorter evolutionary time-period.

Diagram

Description automatically generatedChart

Description automatically generatedFigure 6 also indicates genetic-distance decreases with spatial proximity. This suggests more genetic divergence derives from regional adaptation than halotolerance; SOL/SFG2 possesses the largest saline-contrast, but smallest separating genetic and geographic distances - implying more importance to distance than salinity. SplitsTree analysis in Figure 7 demonstrates most significant interpopulation-pair gene-flow exists between PAU/LLAN and SOL/SFG2, although overall rates of genetic exchange are highly limited. Relatively lower rates of BRUC/GAR interpopulation-pair gene-flow are explained by larger distance from its neighbouring coastal-inland pair. In summary, this pattern indicates rare adaptative events in each coastal location combine with low but existing gene-flow within, and sporadically between, coastal-inland population-pairs that maintains their close relationship - BRUC/GAR represents both the most spatially and phylogenetically-distant pair within members and between populations, likely because both intra- and interpopulation-pair gene-flow is limited by increased physical separation.

We therefore conclude phenotypic contrast is most appropriately analysed through independent pairwise-comparison of three coastal/inland population-pairs distinct to particular Spanish locations - BRUC/GAR, PAU/LLAN and SOL/SFG2. This approach introduces confounding variation; each pair undergoes local adaptation to its specific environment, particularly plausible considering low rates of interpopulation gene-flow. Each *B.fruticulosa* strain only comprises five individuals, therefore each pairwise-comparison will include substantial noise - a significant limitation of this project. However, these shortcomings can be mitigated by ultimately comparing identified variant-outliers across all three population-pairs.

Graphical user interface, diagram

Description automatically generatedTo conduct selection-scans, we employ window-based SNP comparisons between members of each coastal-inland population-pair using the 65Mb assembly subset containing 1,909,722 biallelic-SNPs. This reduced dataset extracts the thirty-nine longest contigs (all >1MB) from the original low-quality assembly, which possesses poor contiguity with an N50 of 121-kilobases and only assembles 82% of the complete genome. This alleviates problems associated with discontiguity whilst minimising computational requirements for analysis. Identifying footprints of differentiation using window-based approaches works most effectively with strong phenotypic-contrast, as exhibited here. Moreover, the characterised recency of coastal colonisation demonstrates signal from haloadaptive-polymorphism is less likely to be obfuscated by noise from additional functional differences arising through evolution. Figure 8 displays adjacent SNP-linkage decays sharply with physical distance for *B.fruticulosa*, therefore window-based approaches represent a conservative technique for identifying genomic variation. Contig-3 represents an exception, as LD-values increase with genetic-separation beyond 30,000-bases, however this could be accredited to chance as few SNPs exist across these distances. High background genome-wide LD likely derives from low nucleotide-diversity between populations.

Chart

Description automatically generated with medium confidenceChart, histogram

Description automatically generatedCombining VCFTools software with our custom R-script, we identified the upper 99th-percentile of fixed 5,000-base windows containing at least 25-SNPs from the empirical FST-distribution for each population-pair. This generated inclusive lists of gene-coding loci displaying significant differentiation within each pair. We used 25-SNP windows to exclude potential FST-estimation bias caused by low-information windows. Window length was fixed to homogenise gene positions across all population-pairs, enabling interpopulation comparison between candidate outlier-windows. 5,000-base window-sizes account for linkage-disequilibrium, which decays on-average by 71% at this distance, whilst isolating only one gene per-window - the average gene-size in relative *Brassica napus* is 2.73-kilobases[14]. Figures 9a-c also demonstrate correlation between FST and SNP-frequency per-window is weak (Spearman’s rank correlation-coefficient ranges from 0.016-0.10 across population-pairs). Incidentally, correlation between FST and SNPs per-window is at least 4x-fold larger for the most genetically-different population (BRUC/GAR), indicating our approach is most effective under low variation. Retrospective studies should stipulate a greater number of SNPs per-window, particularly when analysing more significantly disparate members within pairs, to reduce probability of identifying ‘false-positives’.

Text

Description automatically generated

Chart, histogram

Description automatically generatedCandidate-gene lists totalled 64,38 and 48 for BRUC/GAR, PAU/LLAN, and SOL/SFG2 coastal-inland pairs respectively (Appendices 4-6). Identified candidates require cautious interpretation; lists assume directional-selection dictates loci with greatest differentiation, particularly uncertain considering each pair inhibits a distinct location within Catalonia and may experience local adaptation. However, FST-distributions for each population shown in Figures 10a-c demonstrate most alleles are not under significant variation between coastal-inland partners, indicating historical genetic drift is low and therefore only selective advantageous genes are likely to appear as outliers. These graphs reveal the outlying FST value cut-off is lower in SOL/SFG2 than other spatially-proximate pairs, indicating fewer windows containing 25-SNPs frequently diverge between SOL and SFG2. This supports there are fewer genetic differences between SOL and SFG2, as suggested in our demographic analysis, and demonstrates SOL/SFG2 residuals may have lower frequencies than outliers from other pairs.

Chart, histogram

Description automatically generated

Text

Description automatically generated

Figure 11 demonstrates particular contigs contain different outlier proportions than expected by chance - for example, contigs-17 and 27 possess a significantly greater percentage of outliers, whilst contigs-1 and 8 have significantly less (see Appendix 7). Adaptive-trait complexes are preserved by tight-linkage between co-adapted loci - correspondingly, contigs containing excess outliers present lower linkage-disequilibrium decay-rates and are therefore more likely to mediate biological adaptation (Figure 5). Alleles under selection are also likely to exhibit closer linkage through intraspecies gene-flow[15]. This suggests 5,000-base windows are insufficient to completely mitigate influence of genetic-hitchhiking within candidate-gene lists - linkage-disequilibrium decay-rates are slower in contigs facilitating halotolerance than previously expected, and retrospective studies should increase window-size.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Text

Description automatically generated

Moreover, as shown in Figures 12a-b,GO-analysis of candidate-gene lists combined from every population-pair displays above 1.5-fold enrichment for categories associated with halotolerance, supporting relevance of these candidates to coastal adaptation (Fisher’s exact-test; p < 0.01, p < 0.05 for molecular-function and biological-process categories respectively). Three of the most significantly enriched molecular-functions surround the cytoskeleton and constituent microtubules. Previous studies demonstrate *Arabidopsis* halotolerance requires cortical-microtubule reorganisation to mitigate salt-stress-induced tubulin-depolymerization and enable plant survival[16]. This implicates candidates mediating microtubule activity in parallel adaptation, therefore we further interrogate the three largest pairwise-FST outliers from our gene-lists assigned to corresponding GO-terms, as shown in Figure 13.

Chart, bar chart

Description automatically generated

Chart, histogram

Description automatically generated

Text

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Text

Description automatically generated

Chart, scatter chart

Description automatically generatedThese candidates encode a p-loop-containing NTP-hydrolase superfamily-protein(AT3G12020), an ATP-binding microtubule family-protein(AT5G23910) and microtubule-associated protein 65-8([AT1G27920](https://www.arabidopsis.org/servlets/TairObject?type=locus&name=AT1G27920)), with weighted FST-values of 0.92,0.76 and 0.67 respectively. Interestingly, these genes are observed in PAU/LLAN, BRUC/GAR and SOL/SFG2 pairs respectively, implying consistent method of adaptation. Moreover, all three genes have well-characterised relevance to mediating plant survival in salt-stress. ATP-binding microtubule proteins are requisite for tubulin-elongation, whilst MAP65-upregulation through the phospholipase-D pathway is well-characterised to enhance microtubule-polymerization during salt-stress - cytoskeletal structure is otherwise disrupted by high sodium-concentrations, which can cause plant death through multiple mechanisms including increased pathogen-susceptibility[17,18]. NTP-hydrolase-dependent nitric-oxide production mediates halotolerance by alleviating germination inhibition - which accordingly represents an enriched biological-process[19]. Hydrolases exhibit 1.2-fold coastal enrichment across all three population-pairs, indicating associated upregulation of ROS-scavenging genes represents common salt-stress response within *B.fruticulosa*[20]. We therefore further investigate these three candidates through AFD-plots displayed in Figure 14. Consistently high AFD-values indicate one major mechanism of halotolerance onset in *B.fruticulosa* involves altered microtubule activity, namely through increased tubulin-polymerisation, that maintains cytoskeletal structure under salt-stress.

A picture containing text

Description automatically generated

Scatter chart

Description automatically generated with medium confidence



Chart, histogram

Description automatically generatedGO-analysis also indicates significant fold-enrichment of multiple genes encoding transferases, including pentosyl- and acetyltransferases. These proteins have well-characterised function in synthesis of phytohormones secreted in response to salt-stress[21]. This implicates candidates mediating transferase activity in parallel-adaptation, therefore we again interrogate the three largest pairwise-FST outliers assigned to these GO-terms (shown in Figure 15). Two candidates derive from the PAU/LLAN pair - encoding glucuronoxylan 4-O-methyltransferase-like DUF579(AT1G09610) and PI3K(AT4G29380), with weighted FST values of 0.91 and 0.86 respectively. PAU/LLAN again provides the most differentiated candidate; reduced sample size potentially elevates genetic-differentiation estimates, which remain sufficient to generate overall mechanistic understanding. DUF579 is required for 4‐O–methylation of glucuronic-acid within AGPs, essential cell-wall components that mitigate salt-stress by loosening pectic networks to facilitate cell-expansion under strong osmotic pressures[22]. PI3K-dependent plasma-membrane endocytosis mediates salt-induced ROS-production, associated with increased rates of salt-stress survival[23]. Both genes are therefore implicated in halotolerance onset within *B.fruticulosa*.

Chart, histogram

Description automatically generated

Text

Description automatically generated

Table

Description automatically generatedOur third interrogated gene originates from BRUC/GAR with an FST value of 0.88, encoding glutamine-phosphoribosyl-pyrophosphate amidotransferase-3(AT4G38880). AFD-plots for all three genes are shown in Figure 16, with once again correspondingly high AFD-values. AT4G38880 is upregulated during osmotic stress to increase *de-novo* purine nucleotide synthesis, which ultimately allows more ATP-generation for metabolic processes enabling salt-stress acclimation, such as active sodium efflux[24]. Correspondingly, we observe significant enrichment within biological processes associated with nucleotide synthesis, as well as fatty-acid metabolism, which can also be co-opted for ATP-synthesis through aerobic respiration[25]. We hypothesise increased ATP requirements are linked to increased microtubule polymerisation, an ATP-induced process. Elevated synthesis of ATP-generating substrates therefore represents another mechanism of haloadaptation in *B.fruticulosa*.

Text

Description automatically generatedA picture containing scatter chart

Description automatically generatedA picture containing text

Description automatically generated

A picture containing scatter chart

Description automatically generatedDespite an increased p-value stringency, the biological-process dataset contains many more enriched categories, including circadian-rhythm. Circadian cycles change to coordinate internal-responses with external environmental-stresses - in *Arabidopsis*, GI clock-proteins associated with SOS-complexes experience proteasomal-degradation in high-salinity, liberating SOS1 Na+/H+-antiporters to facilitate sodium efflux[26]. Temporal control of salt-stress response is particularly relevant for coastal populations experiencing rhythmic tidal-patterns. Clock-protein-mediated post-translational regulation may also explain observed RNA-processing variation; modified RNAs are required to adjust clock-protein activity. Candidate-lists contain three genes encoding clock-proteins, one from each population-pair. The largest FST value again derives from PAU/LLAN, encoding cryptochrome-2(AT1G04400) - experimental evidence demonstrates this gene is significantly upregulated under high-saline conditions, and knock-out prompts shortened circadian periods uncoordinated with salinity peaks[27]. The remaining two genes encode histone-deacetylase subunits (AT1G19330) and transcription-factor TCP7(AT5G23280), both implicated in regulating salt-stress response genes from independent feedback-pathways within the circadian-clock[28,29]. AFD-plots for all three genes are displayed in Figure 17. We accordingly speculate altered circadian-rhythm precipitates halotolerance through population-specific mechanisms linking salt-stress response to sodium peaks.

Text

Description automatically generated with medium confidence

Text

Description automatically generatedScatter chart

Description automatically generated

Chart

Description automatically generatedTo further refine divergence-outliers associated with coastal adaptation, we overlapped our three gene-lists to identify candidates implicated in parallel differentiation for all population-pairs. However, only three outliers remain present in at least two populations (shown in Table 3). Nevertheless, the fraction of common gene-candidates (0.02) is consistent with comparably divergent systems exhibiting parallel adaptation[30,31]. Unsurprisingly, conserved outliers constitute heavily enriched GO-categories implicated in halotolerance, namely nucleotide-metabolism, the chitinolytic system and cytoskeletal structure. Indeed, the chitinase family-protein identified in both SOL/SFG2 and BRUC/GAR is induced by salt-stress in *A.thaliana*, and represents focus of efforts to engineer biotic-stress tolerance through eased cellular-expansion[32]. Myosin heavy-chain proteins are also essential for halotolerance, functioning as actin-based motors that drive cytoskeletal translocation of salt-stress response molecules[33]. Figure 18 displays AFD-plots for all three candidates.

Text

Description automatically generated

Text

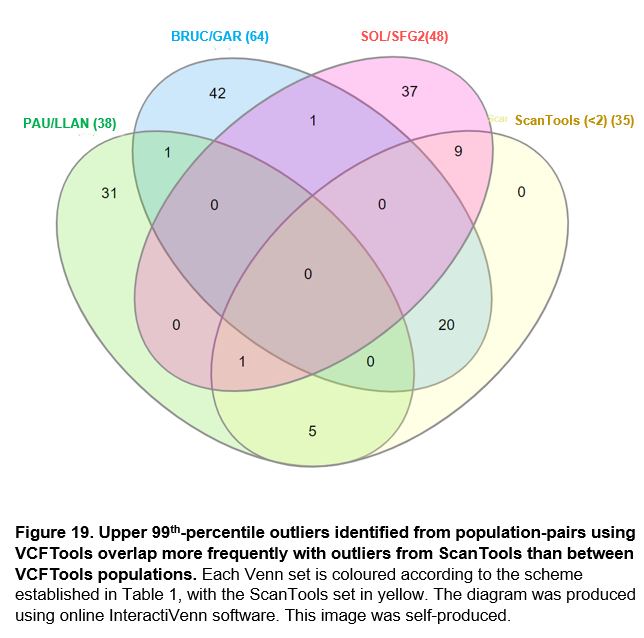
Description automatically generatedA picture containing table

Description automatically generatedTimeline

Description automatically generatedText, letter

Description automatically generated

We therefore applied a second technique - the ScanTools pipeline - to identify excessively-differentiated genes between paired populations, only retaining gene-coding loci observed using both approaches. Here, we again generate candidate-lists of 1%-outliers across 5,000-base windows containing minimum 25-SNPs - a total of 93 candidate-genes are present in at least two population-pairs (Appendix 8). Figure 19 demonstrates if we compare ScanTools outliers with those identified in at least one population with VCFTools, we detect a much higher frequency of shared outliers(35). Indeed, conserved outliers are significantly more frequent than expected by chance alone (38 vs 1.17). Candidates also present consistent enrichment in GO-categories relevant to saline-adaptation including transferase activity and nucleotide-metabolism, as displayed in Figure 20 (Fisher’s exact-test; p < 0.05 for both categories).



Chart, bar chart

Description automatically generated

Text

Description automatically generated



**B**



Our ScanTools-based approach, which concatenates candidates from multiple population-pairs, reduces false-positives from genetic drift and population-specific selection, but limits our ability to identify global pathways altered by different genes. This represents a significant weakness, as the absence of common candidates between all three VCFTool-pairs and minimum two pairs from ScanTools implies a complex, multidimensional genetic-basis for halotolerance, supported by significant functional parallelism between enriched-candidates. Soft-sweep adaptation may also reduce selection-scan power, however limited standing genetic-variation between population-pairs renders this unlikely. Ultimately, we identify significant enrichment discerning particularly robust outliers instigating functional consequence. However, the inability of FST to account for opposing effects of genetic drift and gene-flow means implementing a second metric such as DDR would increase reliability of proposed mechanisms for halotolerance onset[34]. Diversity/Differentiation-Residual graphs and identified outlier-windows are found in Appendices 9-12. One overarching limitation from both metrics is SNP-data only represents one component of the genetic landscape, and despite linkage, fails to measure complete structural-variation. Transposon-element and/or transcriptome data is required to further elucidate genomic mechanisms underlying parallel-haloadaptation, ultimately rationalising crop design with improved abilities to colonise agriculturally-challenging environments.

**Appendices:**

Appendix 1 = *B.fruticulosa* annotation file.

Appendix 2 = *B.fruticulosa* orientation file

Appendix 3 = Isolation-By-Distance Coordinates

Appendix 4 = BRUC/GAR outliers

Appendix 5 = PAU/LLAN outliers

Appendix 6 = SOL/SFG2 outliers

Appendix 7 = Number of outliers per contig (actual versus expected)

Appendix 8 = Conserved outliers between population pairs

Appendix 9 = DDR graphs for each population pair

Appendix 10 = Top 1% DDR outlier windows for BRUC/GAR

Appendix 11 = Top 1% DDR outlier windows for PAU/LLAN

Appendix 12 = Top 1% DDR outlier windows for SOL/SFG2

**Word Count:** 3431

**References:**

1. Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., ... & 1000 Genomes Project Analysis Group. (2011). The variant call format and VCFtools. *Bioinformatics,* **27(15):** 2156-2158.

2. Ruan, J. & Li H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat Methods,* **17(2):**1 55–8.

3. Chen, J. P., Ge, X. H., Yao, X. C., Feng, Y. H., & Li, X. Y. (2011). Synthesis and characterization of interspecific trigenomic hybrids and allohexaploids between three cultivated *Brassica* allotetraploids and wild species *Brassica fruticulosa*. *African Journal of Biotechnology,* **10(57):**12171-12176.

4. Gupta, B., & Huang, B. (2014). Mechanism of salinity tolerance in plants: physiological, biochemical, and molecular characterization. *International journal of genomics,* **2014:** 701596.

5. Chevilly, S., Dolz-Edo, L., Morcillo, L., Vilagrosa, A., López-Nicolás, J. M., Yenush, L., & Mulet, J. M. (2021). Identification of distinctive physiological and molecular responses to salt stress among tolerant and sensitive cultivars of broccoli (*Brassica oleracea* var. *Italica*). *BMC plant biology,* **21(1):** 488.

6. Huson, D. & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Mol. Biol. Evol.,* **23(2):**254-267.

7. Via, S. (2012). Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Biological sciences,* **367(1587):** 451–460.

8. Jombart, T (2008). Adegenet: an R package for the multivariate analysis of genetic markers. *Bioinformatics,* **24:** 1403-1405.

9. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., ... & Durbin, R. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics,* **25(16):** 2078-2079.

10. Chen, Y., Ye, W., Zhang, Y., & Xu, Y. (2015). High speed BLASTN: an accelerated MegaBLAST search tool*. Nucleic acids research,* **43(16):** 7762-7768.

11. Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., ... & Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome research,* **13(9):** 2129-2141.

12. Heberle, H., Meirelles, G. V., da Silva, F. R., Telles, G. P., & Minghim, R. (2015). InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. *BMC bioinformatics,* **16(1):** 1-7.

13. Catalonia Region Profile. (2018, June 11). BBC News. Retrieved from <https://www.bbc.co.uk/news/world-europe-20345071>

14. Chen, X., Tong, C., Zhang, X., Song, A., Hu, M., Dong, W., ... & Zhang, L. (2021). A high‐quality *Brassica napus* genome reveals expansion of transposable elements, subgenome evolution and disease resistance. *Plant biotechnology journal,* **19(3):** 615-630.

15. Bürger, R. & Akerman, A. (2011). The effects of linkage and gene flow on local adaptation: a two-locus continent-island model. *Theoretical population biology,* **80(4):** 272–288.

16. Wang, C., Li, J. & Yuan, M. (2007). Salt tolerance requires cortical microtubule reorganization in *Arabidopsis. Plant & cell physiology,* **48(11):** 1534–1547.

17. Valiron O. (2011). New insights into microtubule elongation mechanisms. *Communicative & integrative biology,* **4(1):** 10–13.

18. Quentin, M., Baurès, I., Hoefle, C., Caillaud, M. C., Allasia, V., Panabières, F., Abad, P., Hückelhoven, R., Keller, H., & Favery, B. (2016). The *Arabidopsis* microtubule-associated protein MAP65-3 supports infection by filamentous biotrophic pathogens by down-regulating salicylic acid-dependent defenses. *Journal of experimental botany,* **67(6):** 1731–1743.

19. Zhao, M. G., Tian, Q. Y., & Zhang, W. H. (2007). Nitric oxide synthase-dependent nitric oxide production is associated with salt tolerance in *Arabidopsis. Plant physiology,* **144(1):** 206–217.

20. Zhou, Z., Zhi, T., Han, C., Peng, Z., Wang, R., Tong, J., Zhu, Q., & Ren, C. (2020). Cell death resulted from loss of fumarylacetoacetate hydrolase in *Arabidopsis* is related to phytohormone jasmonate but not salicylic acid. *Scientific reports,* **10(1):** 13714.

21. Verma, V., Ravindran, P., & Kumar, P. P. (2016). Plant hormone-mediated regulation of stress responses. *BMC plant biology,* **16:** 86.

22. Temple, H., Mortimer, J. C., Tryfona, T., Yu, X., Lopez-Hernandez, F., Sorieul, M., Anders, N., & Dupree, P. (2019). Two members of the DUF579 family are responsible for arabinogalactan methylation in *Arabidopsis.* *Plant direct,* **3(2):** e00117

23. Leshem, Y., Seri, L. & Levine, A. (2007). Induction of phosphatidylinositol 3-kinase-mediated endocytosis by salt stress leads to intracellular production of reactive oxygen species and salt tolerance. *The Plant journal: for cell and molecular biology*, **51(2):** 185–197.

24. Massange-Sánchez, J. A., Palmeros-Suárez, P. A., & Délano-Frier, J. P. (2016). Overexpression of Grain Amaranth (*Amaranthus hypochondriacus*) AhERF or AhDOF Transcription Factors in *Arabidopsis thaliana* Increases Water Deficit- and Salt-Stress Tolerance, Respectively, via Contrasting Stress-Amelioration Mechanisms. *PloS one,* **11(10):** e0164280.

25. Chen, X., Zhang, L., Miao, X., Hu, X., Nan, S., Wang, J., & Fu, H. (2018). Effect of salt stress on fatty acid and α-tocopherol metabolism in two desert shrub species. *Planta,* **247(2):** 499–511.

26. Park, H. J., Kim, W. Y., & Yun, D. J. (2013). A role for GIGANTEA: keeping the balance between flowering and salinity stress tolerance. *Plant signaling & behavior,* **8(7):** e24820.

27. Doi, M., Takahashi, Y., Komatsu, R., Yamazaki, F., Yamada, H., Haraguchi, S., Emoto, N., Okuno, Y., Tsujimoto, G., Kanematsu, A., Ogawa, O., Todo, T., Tsutsui, K., van der Horst, G. T., & Okamura, H. (2010). Salt-sensitive hypertension in circadian clock-deficient Cry-null mice involves dysregulated adrenal Hsd3b6. *Nature medicine,* **16(1):** 67–74.

28. Zheng, Y., Ding, Y., Sun, X., Xie, S., Wang, D., Liu, X., Su, L., Wei, W., Pan, L., & Zhou, D.X. (2016). Histone deacetylase HDA9 negatively regulates salt and drought stress responsiveness in *Arabidopsis*. *Journal of experimental botany,* **67(6):**1703–1713.

29. İlhan, E., Büyük, İ., & İnal, B. (2018). Transcriptome - Scale characterization of salt responsive bean TCP transcription factors. *Gene,* **642:** 64–73.

30. Takuno, S. *et al.* (2015). Independent molecular basis of convergent highland adaptation in maize. *Genetics,* **200:** 1297–1312.

31. Lai, Y. T. *et al.* (2019). Standing genetic variation as the predominant source for adaptation of a songbird. *Proc. Natl Acad. Sci. USA,* **116:** 2152–2157.

32. Takenaka, Y., Nakano, S., Tamoi, M., Sakuda, S., & Fukamizo, T. (2009). Chitinase gene expression in response to environmental stresses in *Arabidopsis thaliana*: chitinase inhibitor allosamidin enhances stress tolerance. *Bioscience, biotechnology, and biochemistry*, **73(5):** 1066–1071.

33. Kosová, K., Práil, I. T. & Vítámvás, P. (2013). Protein contribution to plant salinity response and tolerance acquisition. *International journal of molecular sciences,* ***14*(4):** 6757–6789.

34. Flanagan, S. P., & Jones, A. G. (2017). Constraints on the FST-Heterozygosity Outlier Approach. *The Journal of heredity,* **108(5):** 561–573.