# Genome-wide cline analysis identifies new locus contributing to a barrier to gene flow across an *Antirrhinum* hybrid zone

**Authors**: Field, David L\*, Coen, Enrico, Barton, Nicolas H… & others

**Affiliation**: IST Austria, Am Campus 1, 3400 Klosterneuburg, Austria

**Running** **title**: Genome-wide clines in snapdragons

**Keywords**: Genome scans, FST, hybrid zones, clines, selection, genomic clines, flower colour

* Re-do genome scans ref v3.5 & update slidingWindows Python 3+ (David)
* Re-do clines ref v3.5 & update fastClines v1.4 to Python 3+ & sync files (David)
* Recreate all figures where required (David)
* Insert RUBIA section, foreshadow in intro, discussion (David & Des/Rico, Sean)
* Permutation tests & Anthocyanin pathway genes (Sean, Dasha, David)
* Updating Introduction & Discussion (Sean)
* Update anything relevant from Tavares (ROSel) & Bradley et al (Sulf)
* Removing repetition from tavares Fst scans, SNP methods etc.
* Significance tests, more cautious considering issue of independence
* Add better connection with flower colour literature Intro (Sean)
* Cutting recombination parts (David)
* Revising the 4pq method details (Nick)
* Dealing with overpass, clarifying in text
* Fst in Fig 9 and Fig S9
* SNP genotypes KASP for all individuals (Parvathy)
* Discuss issue of 4pq, different centre or asymmetric tails? (Nick, David, Pavarthy)
* Power simulation section (David)

# ***Abstract – New Feb 2019 following Rico meeting***

# Identification of the genomic regions that contribute to reproductive isolation is a major goal of evolutionary genetics. Much effort has focused on locating candidate genes by scanning genomes for regions of excess FST. However, it is unclear whether all barrier loci will be identified from genome scans. Here we develop a complementary approach based on genome wide screening for geographic clines. We apply this method to a hybrid zone between *Antirrhinum* populations with different flower colour patterns. **Tractable system - Yardstick**. Most genomic regions with steep clines correspond to prominent islands of divergence containing previously identified flower colour loci. However, cline analysis identified an additional locus controlling magenta intensity that was less evident from Fst scans alone. We show this locus tags a gene RUBIA, which interacts with other loci to control flower colour pattern. By comparing features of cline analysis and Fst, we show the strength and weaknesses of each approach and how in combination they provide a more complete and robust evaluation of the genomic regions contributing to reproductive barriers.

Notes from meeting with Rico regarding proposed new structure to manuscript

# ***Introduction***

# ***Results***

## *Genome wide cline analysis identifies additional loci*

## *RUBIA locus controls magenta intensity*

## *RUBIA likely corresponds to UDP glucose*

## *Comparison between cline analysis and Fst*

# ***Conclusions***

## *Relative merits of Fst and cline analysis*

# ***Methods***

**Abstract – OLD version from 2016 manuscript**

Identification of the genomic regions that contribute to adaptation and reproductive isolation is a major goal of evolutionary genetics. Much effort has focused on locating candidate isolating genes by scanning genomes for regions of excess FST. However, such patterns can be generated by many different processes. Scanning genomes for steep clines provides an alternative approach to identify selected genes, However, it may be difficult to distinguish these from the neutral genomic background. In this study, we examine whole genomes from a hybrid zone between two subspecies of snapdragon, *Antirrhinum* *majus pseudomajus* (magenta flowers) and *A. majus. striatum* (yellow flowers). Differences in flower colour are controlled primarily by interactions of a few loci of major effect. We examine how narrow clines and genomic divergence (FST) are distributed in relation to flower colour differences. Regions of elevated FST are strongly associated with the primary colour genes. For some regions, elevated FST was driven by reduced polymorphism associated with reduced recombination rates. The majority of steep clines across the genome are distributed in close proximity to regulatory or structural genes involved in flower colour. We find that the loci with the steepest geographic and genomic clines coincide with those within or tightly linked to regulatory genes involved in magenta pigmentation. Clines near loci involved in yellow pigmentation were more numerous, but shallower and indistinguishable from the genomic average, This result may reflect weaker selection at this locus and enhanced hitchhiking facilitated by the lower recombination rate in this genomic region. Steep clines generally coincide with regions of highest excess FST, but many other divergent regions posses no detectable clines. Taken together, this provides a crucial test for reliability of using genome scans of cline parameters for detecting loci known to be under selection.

*Introduction*

Identifying the genes involved in reproductive isolation and local adaptation, and finding how they are distributed across the genomes, may provide insight into the processes that drive divergence and speciation. Genome-wide scans between divergent populations are commonly used to identify candidate loci involved in local adaptation and reproductive isolation. Regions of excess relative divergence (measured by FST) are often termed ‘genomic islands’, and are often interpreted as resulting from selection on one or more loci that causes a barrier to introgression of linked neutral variation. Thus preserving divergence after the rest of the genome has been homogenized by gene flow (Wu 2001; Turner *et al.* 2005). However, it is still unclear what causes genomic islands, as these signals may also be produced by reduced variation within the parental populations, without any involvement of gene flow or reproductive isolation (Charlesworth 1998; Cruickshank & Hahn 2014). Moreover, given that patterns of diversity and divergence typically reflect long-term evolutionary processes, they may tell us little about current selection or barriers to gene flow across genomes.

Another way to find genes under divergent selection is to map steep clines in naturally occurring hybrid zones. In such zones, genomes mix to produce new gene combinations. If these are less fit, either inherently or because they are in an unfavorable environment, then a stable equilibrium may be reached between admixture and selection (Haldane 1948; Slatkin 1973; Barton & Hewitt 1985). Divergently selected alleles are expected to resist introgression, whereas neutral or advantageous alleles will exchange freely across the hybrid zone. The strength of geographic cline analysis is that local equilibrium is reached quickly for selected (*t*=1/*s*) and neutral loci (*t*=x2 for distance x) (Barton & Hewitt 1985; Barton & Gale 1993). Moreover, when the dispersal rate is known, the strength of selection at individual loci can be estimated (Haldane 1948; Endler 1977; Barton & Gale 1993). Therefore, given enough generations since secondary contact, loci under divergent selection can be located because they display steeper gradients in allele frequencies as a function of distance (*geographic* *clines*) or hybrid index (*genomic clines*; Szymura & Barton 1986; Gompert & Buerkle 2010) compared to neutral markers. Assessing cline centre across genomes is also useful, because epistasis and differences in the geographic location of selection are expected to shift cline centres (Gavrilets 1997). However, theory predicts that local population structure (physical barriers) and linkage disequilibrium tends to pull clines together (Barton 1986), whilst drift scatters clines (Polechová & Barton 2011), obscuring these simple expectations. Nevertheless, when combined with knowledge of the functional genes at these loci and how they relate to phenotypic variation and reproductive isolation, the presence of steep clines provides strong evidence for the presence of a genetic barrier.

In contrast to the presence of steep clines, regions of elevated FST may arise by many different processes unrelated to current contact between populations. Given FST is a relative measure of genetic differentiation , either reduced diversity within (*w*) or increased sequence divergence between populations (*b;i.e. D*xy) can increase FST. As a consequence, elevated FST may arise due to some combination of processes such as selective sweeps (in allopatry), background selection (Charlesworth *et al.* 1993), stochastic effects of drift and reduced diversity in regions of low recombination (Noor & Bennett 2009). These processes may occur in the complete absence of gene flow and thus not directly related to heterogeneous rates of gene flow across the genome (Cruickshank & Hahn 2014). Therefore in hybrid zones where gene flow is ongoing, examining patterns of clines across the genome together with FST can help disentangle patterns related to contemporary selection from historical processes that shaped genome wide divergence.

There is a long history of using cline theory at small numbers of loci to estimate selection in hybrid zones (Haldane 1948; Endler 1977). With genome wide data, we are interested in both the distribution and frequency of steep clines in different regions of the genome as well as their individual properties including width and centre (position). This raises new statistical challenges, especially in distinguishing signals of selection from the stochastic noise across large numbers of loci. The distribution of cline properties will depend on several parameters that may be difficult to measure in nature, including: the time since secondary contact, neighbourhood size and local recombination rate. Thus, interpreting patterns of genome wide clines is challenging, knowing at least some true positives (i.e. genes under divergent selection) allows inferences to be validated. To assess the reliability of detecting selection by scanning whole genomes for steep clines, the location of known isolating genes provides an important reference to compare against the neutral background.

In angiosperms, flower colour provides a tractable model for investigating genetic barriers to gene flow. Transitions in flower colour often have a simple genetic basis, are well characterized at the molecular level and have a clear link with agents of selection and reproductive isolation (e.g. Bradshaw & Schemske 2003; Rausher 2008; Streisfeld *et al.* 2013). In this study, we use whole genome data from across a natural hybrid zone between two subspecies of the snapdragon, *Antirrhinum* *majus* (*A. m.* *pseudomajus* and *A. m.* *striatum*), to examine whether the patterns of clines across the genome is consistent with selection on flower colour genes. These two subspecies are only distinguishable by flower colour, *A.* *m.* *pseudomajus* being predominately magenta in contrast with the yellow *A.* *m.* *striatum* (Fig 1). These phenotypes are determined by the interaction of a few loci that regulate different components of the flavonol biosynthetic pathway and influence the pattern of pigmentation of anthocyanin (magenta) and aurone (yellow) in different parts of the flower. The magenta intensity and patterning across the flower is primarily controlled by two MYB-like transcription factors, *Rosea* (Schwinn *et al.* 2006) and *Eluta*, which are tightly linked (0.5cM apart) (Tavares *et al.* in prep). The genetic basis of yellow pigment variation is less well understood, but several structural genes (AS1, *Am*4’CGT) involved in the aurone branch and expression of yellow pigment have been isolated (Ono *et al.* 2006). Other structural flavonol pathway genes and regulatory genes that influence the intensity and distribution of pigmentation have been also isolated and mapped (see also Schwinn *et al.* 2006). This set of well-characterised flower colour genes, some with a direct connection to phenotypic variation under selection, makes snapdragons a useful model for studying the origins and mechanisms of reproductive isolation.

In a few locations where the two subspecies meet in the Pyrenees in Spain and France, they interbreed and generate recombinant colour phenotypes across narrow hybrid zones. Although previous studies found no evidence for post-zygotic barriers (Andalo *et al.* 2010), observations and manipulative experiments suggest that pollinator (*Bombus* *hortorum*) mediated selection on flower colour causes reproductive isolation (Ellis & Field in prep). Moreover, steep clines in phenotype and allele frequencies at the *Rosea* and *Eluta* genes suggest strong selection at two of the main flower colour genes (Tavares *et al.* in prep; Whibley *et al.* 2006). However, to date, most of what we know about genetic barriers has focused on a small portion of the genome (~1Mb) encompassing *Rosea* and *Eluta* (Tavares *et al.* in prep). Whether steep clines are restricted to regions tightly linked to *Rosea*, *Eluta* and other flower colour genes, or dispersed across the genome is crucial for understanding the genetic architecture of reproductive isolation.

Here, we relate the density and properties of clines to genome wide divergence, recombination rates and proximity to functional flower colour genes. A few studies have assessed associations between FST and genomic cline parameters (Gompert *et al.* 2012; Nosil *et al.* 2012; Parchman *et al.* 2013). However, in these studies, the genomic distribution of the loci examined and their relation to phenotypic traits involved in reproductive isolation was unknown. Here, we use pooled whole genome data and develop a novel approximation method for efficiently estimating cline parameters for large numbers of loci. Taking advantage of the genetic resources available for *Antirrhinum*, we combine the location of the flower colour genes with a draft reference genome and genetic map, and develop a series of predictions based on cline theory, to ask:

1. How is divergence distributed across the genome? Does diversity *within* or *between* populations drive patterns of relative divergence? How does this relate to recombination rates and candidate genes? We use sliding window analyses of diversity and divergence statistics between multiple *pseudomajus* and *striatum* populations.
2. Are clines significantly clustered across the genome? Are steep clines associated with known genes related to reproductive isolation? Do geographic and genomic clines detect similar outlier loci? We estimate cline parameters at individual loci across the genome and examine how these are distributed in relation to genes known to affect flower pigmentation.
3. How do clines and divergence correspond across the genome? We compare patterns of divergence (FST), diversity and and cline (geographic and genomic cline) parameters.

*Methods*

**Whole Genomes**

*Sampling*

In 2013, we sampled plants in a hybrid zone near Planoles, Spain (Fig 1). A total of *n* = 1600 plants were located to within 2 metres with a GPS (Trimble GeoXT datalogger), leaf tissue collected for DNA extraction and one flower taken for phenotyping. Following Whibley et al., (2006), individuals were categorized into six phenotype/genotype classes on the basis of anthocyanin and aurone pigmentation across the flowers. For genomic analyses we randomly selected a subset of 50 unrelated individuals from each of six subpopulations. To ensure spatial coverage of the flower colour cline, three subpopulations were selected in the predominately yellow regions west of the centre of the cline and three in magenta dominated regions to the east (see Fig 1). At each site, the individuals were randomly selected, at least 5 metres apart from one another within a 200 metre radius. In the outermost populations (YP1 and MP6) only yellow and only magenta individuals are present, respectively. However, in the remaining (YP2, YP3, MP1, MP2) hybrids and both parental phenotypes are present (Fig 1). The plants are mostly restricted to within 100 metres of the two roughly parallel roads passing through Vall de Ribes. Altitude gradually increases up the valley going West, above ~1600 metres in altitude, *Antirrhinum* is absent. Thus, a break of ~3 kilometres in the distribution of *Antirrhinum* plants coincides with a mountain pass. The outermost yellow population (YP1) is situated west of this pass, while the other five population samples are east of the mountain pass (Fig 1).

*Whole genome sequencing, alignment and SNP calling*

Whole genome sequencing was carried out by John Innes Genome Centre (Norwich, UK) using Illumina HiSeq and has been described in detail elsewhere (Tavares et al., in prep). For the six subpopulations, DNA was extracted individually and then pooled, with each pool barcoded and sequenced with at least 110bp paired-end reads. For the *pooled* individuals, reads were aligned to a draft reference genome (Yong unpublished) for *Antirrhinum majus* consisting of 6,000 scaffolds and 20,000 contigs (N50 value?). To avoid biases introduced from variable read lengths, all were equally trimmed to yield 100bp paired-end reads. Alignments were performed with Bowtie (Lagmead et al 2009) using the default settings and ambiguous mapped reads were removed with samtools (ref). Scaffolds and SNPs were placed on a draft genetic map generated for *Antirrhinum* based on a cross between the highly inbred *Antirrhinum majus* reference individual and *Antirrhinum molle*. The map positioned 1347 of the largest scaffolds, encompassing 444Mbp (70%) of the ~630Mbp *Antirrhinum* genome (Yong unpublished). Scoring of SNPs employed custom Python scripts (supporting information 1). Sites with less than 20x coverage in all six pools, singletons that occurred in fewer than two pools and sites with more than two bases were excluded from the data set. This yielded mean sequence depths in the outermost pools of and and between and in the central pools.

*Genomic diversity and divergence*

We quantified relative genetic divergence and diversity *within* and *between* the six pooled populations. To avoid bias due to sequencing errors and paralogy, we excluded loci with low (<20) or high coverage (> 200) in any of the populations. Site-specific allele frequencies were estimated in each of the subpopulations for individual loci using allele counts. We estimated diversity within (*w*) each of the populations, and total diversity (*t*), relative (*F*ST) and absolute differentiation (*b;i.e. Dxy*) between each pair of populations. Due to potential biases with pooled sequence data, we made adjustments to the summary statistics to account for sampling errors to *w* and *t* (Futschik & Schlötterer 2010). We implemented a sliding window analysis of these summary statistics for various window sizes (5, 10, 50, 100kB) and overlap sizes (2.5, 5, 25, 50kB, respectively). We calculated relative differentiation for each window as , where The windowed mean values of diversity and relative differentiation were used to calculate absolute differentiation (i.e. pairwise sequence divergence), which can be approximated as *b* = (1+*F*ST)/(1-*F*ST). Windows were excluded when coverage (proportion of window with minimum read depth < 20) was below 0.1. All summary statistics and sliding window analyses were conducted in a custom Python script (need to archive this package). Here, we only report 10kB because smaller resulted in excessive variance and larger windows less precision in relation to candidate genes.

*Geographic clines*

Estimating geographic cline parameters for whole-genome data is not straightforward, especially when alleles are not fixed for alternative alleles in the most outer populations (when many parameters need to be fitted for thousands of loci across whole genomes. One approach is the cline approximation method described by Polechova & Barton (2011), which is not computationally demanding and provides a reasonable approximation to detailed cline model fitting (e.g. Szymura & Barton 1986; Gay *et al.* 2008). They showed that the width of a cline can be approximated for *diagnostic loci* as the integral of heterozygote frequencies over space. For discrete demes, this is essentially twice the sum in the frequency of heterozygotes across all demes,

where, *nd* is the number of demes, *pi* and *qi* the allele frequencies in the *i*th deme, and width is measured in deme spacing. Similarly, the centre of the cline can be estimated as the sum of allele frequencies adjusted by half a deme,

However, with real data loci may not be fixed for alternative alleles in one or both parental populations. Moreover, this model assumes equal spacing among demes.

We extended this approach to non-diagnostic loci that also accounts for differences in deme spacing across a hybrid zone. For diagnostic or non-diagnostic loci, we denote the allele frequencies in the two parental populations (or outer most demes) as *p*0 and *p*1 and assume that they are known. We then restrict the integral (or sum) to within the interval of the allele frequency differences in the parental populations. The cline width can then be approximated by,

where *di* is the span of the deme, and equal to half the Euclidean distance between the midpoint of the samples in each deme. For the outermost demes we extend this outwards to the same distance. This term *di* provides the appropriate scaling to account for irregular spacing of demes. Similarly, cline centre is approximated by,

where the first term orientates the centre with respect to *p*0.

We use these methods for the *Antirrhinum* data, taking the pooled allele frequencies as given across the hybrid zone. Here the demes were irregularly spaced apart and we scaled according to the midpoint distance between the demes [*di* = (6000,5000,1500,1400,3500,6000)]*.* Although allele frequencies estimate may benefit from incorporating errors for pooled data (e.g. Lynch *et al.* 2014), we found a high correlation between allele frequencies estimated from the pools and those estimated from individual genotyping of the same 50 individuals in each pool (*r*2 =xx; Supporting Information Sxx). We include only loci with strong allele frequency differences between the outer pools ≥ 0.80 and ≥ 0.90. For these loci, allele frequencies were polarized so that they increased from West to East (yellow to magenta, respectively). One limitation of this method is that cline reversals (e.g. *p*5 *> p*6) for non-diagnostic loci can result in negative cline widths. However, we only identified three such loci with significant reversals. Considering this occurred at loci with significantly lower depth in the parental populations than the average, this may be due to sampling artifacts and therefore we removed these from further analyses.

To test for mixed distributions in cline parameters, we used Hartigans' dip test for unimodality/multimodality with simulated *P*-value (based on 10000 replicates). For geographic clines, there is no formal way to test whether a particular loci exhibits parameters considered as an outlier compared to the genome wide average. Therefore, we simply take the upper 95% quantiles of cline widths to highlight loci with narrower clines than the bulk of the genome. With course spatial sampling (i.e. six pools), width and centre can be confounded, resulting in narrower clines as the centres move towards the edges of population samples. This is particularly evident at loci with implied narrow clines between YP1 and YP2 (6-7km), often reflecting fixed differences between these populations in the absence intervening plants. Considering this represents strong allele frequency differences but not the presence of a sigmoid cline, we excluded these loci from qualitative comparisons and quantile calculations. Given this method provides only point estimates, we use qualitative differences for comparing loci across the genome. Preliminary checks against simulations that incorporate sources of error (sampling and sequencing) indicate this method provides a reasonable approximation of cline parameters (Supporting Information S1), but this will be investigated in more detail elsewhere. All estimates were calculated in a custom Python script for cline approximations for whole genome data (need to archive this package).

*Genomic clines*

An alternative way to evaluate variation among loci is to use non-geographic cline methods (i.e. genomic clines). This approach, first described by Szymura & Barton (1986) fits clines (allele frequencies) against the mean hybrid index, instead of geographic distance. These methods are especially useful where the geographic pattern of hybridization is complex as in mosaic hybrid zones, or when some loci strongly deviate from the others. Other genomic cline methods instead fit individual diploid genotypes against the mean hybrid index (Gompert & Buerkle 2010), but are not appropriate here given our pooled genomic data only provides allele frequencies. We use the logit-logistic approach (Fitzpatrick 2013) to describe the expected allele frequency for a given locus in terms of theexpected mean hybrid index across all loci, *S* as:

where *u* is the relative difference in cline position (centre) and *v* gives the relative difference in cline gradient (width) compared to the genome wide average. Perfect concordance between a locus and the genome-wide mean hybrid index ( = *S*) when We assume the joint distribution of parameters follows a multivariate normal distribution and use the squared Mahalanobis distance *D*2 of each locus from the mean to detect outliers. Following a distribution with degrees of freedom equal to the number of parameters, we define critical values from Bonferroni adjusted *P* values. Analysis was restricted to loci with strong allele frequency differences (≥ 0.90). To ensure loci showed fixed differences for the calculation of the hybrid index, we introduced parental populations fixed for alternative alleles. We also performed the original concordance method (Szymura & Barton 1986) and obtained very similar results. Genomic cline analysis were performed using custom scripts and *HiEst* (Fitzpatrick 2013).

*Genomic heterogeneity of divergence and clines*

One goal of this study is to determine whether regions with elevated divergence are related to charateristics of clines or other genomic features including recombination rates. Comparative tests of 10kB windows, harbouring steep clines vs. those without clines were made for the most distant pairwise comparison (YP1 x MP6) and those on the same side of the mountain pass (YP2 x MP5). Measures of relative divergence FST, average polymorphism between populations *b* and polymorphism within *w* (for *striatum* and *pseudomajus*) were compared between the two groups of windows (clines present, clines absent) using non-parametric Mann-Whitney U-test.To assess whether genomic divergence correlated with recombination rate, we determined the approximate region of the centromeres based on the relation between physical size and map length. For each chromosome separately, we calculated the distance of each 10 Kb window from the centromere (in cM) and fitted FST using a beta regression model (Cribari-Neto & Zeileis 2010). This model assumes the dependent variable (FST) is beta-distributed with parameterization using mean and a precision parameter (phi = alpha) and assuming a logit link function. Model fitting conducted using *betareg* in R.

*Divergence, clines and gene associations*

Here we examine whether FST and clines detect loci known to be under selection. We use the broad term ‘colour related gene’ to include all of those involved in the flavonol biosynthetic pathway or known to regulate the expression (intensity or distribution) of colour pigments across parts of the flower in *Antirrhinum majus*. Some of these genes have been confirmed to influence phenotypic variation between *pseudomajus* and *striatum* (e.g. MYB-related transcription factors *Rosea* and *Eluta*). However, others have been identified through genetic screens between wild type and mutant lines in *Antirrhinum majus,* but their importance for differences between these subspecies is unknown. These include structural and enzyme producing genes crucial in different parts of flavonol biosynthetic pathway (e.g. FLS, Am4’CGT) as well as other regulatory genes (e.g. *Venosa*). In order to test these associations, we compare estimates of cline parameters for colour genes (and those tightly linked) to the genomic background. Here, we distinguish three sets of loci, those (i) within the structural or regulatory genes that influence colour pigmentation in flowers, (ii) tightly linkedtocolourrelated genes (<200Kb, < 0.5cM), and (iii) background loci (>200Kb or 0.5 cM from known colour related genes). Other loci known to be involved in flower colour in *Antirrhinum* were located from GenBank (see below). Given the tendency for physical linkage between neutral and selected loci to result in hitchhiking and steep clines in adjacent genomic regions, we consider loci <0.5cM of the causal gene to be ‘colour linked’. Outlier loci were detected from geographic and genomic cline methods (see above) to test whether loci within colour related genes exhibit significantly different cline parameter estimates compared to the genomic background.

*SNP genotyping*

We genotyped tissue samples at four SNP loci using KASPR (LGC Genomics) for a larger number individuals (*n* = 1600) across the hybrid zone. These SNPs are a subsample of a panel used for a wider survey of the hybrid zone (David Field, unpublished data). We identified candidate SNPs from whole genome sequence data (Illumina HiSeq) of six pools of 50 individuals across a transect through the hybrid zone. The four loci genotyped include diagnostic or strongly differentiation loci from a representative of the major clusters of cline centres identified from pool seq.

*Results*

From the whole genome sequences, ~6x108 sites aligned to the draft reference genome in all six pools. Of these sites, 2.3 to 3.6% were variable across the pools (after filtering for low depth, singletons and non-biallelic SNPs), yielding 1.3 x 107 to 2.1 x 107 polymorphic loci. Very few fixed differences between the two subspecies were detected (*n*=269 loci) and these were restricted to just two regions of the genome, sites within or tightly linked to *Rosea* on linkage group 6 (LG6) (4 of 269 loci) and *Am*4’CTG on LG2 (265 of 269). There was a substantial difference in the number of loci fixed in one subspecies yet polymorphic in the other. A total of 547 loci were fixed in the outer magenta pool (MP6) and yet polymorphic in the yellow pools, whereas only 30 loci were fixed in the outer yellow pool (YP1) and polymorphic in the magenta pools. In contrast to fixed differences, these loci were located across all linkage groups, but 80% of loci fixed in the outer magenta pool (MP6) and yet polymorphic in the yellow pools were also tightly linked to *Am4’* CTG. This difference was also reflected in the overall higher levels of nucleotide diversity (*w*), which were significantly higher in the far yellow population (YP1,*w*=0.0086) compared to all the other populations sampled (*w*< 0.0075; Mann Whitney U test, *p* < 2.2 x 10-12).

*Genomic islands of divergence*

The pattern of genome-wide divergence (FST) was substantially different among population pairs (Fig 2). Background levels of FST were substantially higher and more variable in the most distant interspecific population pair (YP1 x MP6, median FST = 0.057, SD = 0.062, Fig2A) compared to those in closer proximity (e.g. YP2 x MP5, median FST = 0.037, SD = 0.017, YP3 x MP4, median FST = 0.037, SD = 0.016, Fig2B; see FigS2 for all inter subspecific pair-wise comparisons). Here, the maximal relative divergence for YP1 x MP6 (FST = 0.96), was substantially higher than the closer inter-subspecies comparisons YP2 x MP5 (FST = 0.36) and YP3 x MP4 (FST = 0.44). Surprisingly, the most distant yellow population (YP1) exhibited similar background levels of FST with both magenta and yellow populations (e.g. YP1 x YP2, median FST = 0.053, Fig2C) on the other side of the mountain pass. Thus, excess divergence and significant population structure, irrespective of sub species, is strongly aligned with this geographic feature.

Above the low background level of divergence, regions of excess FST (99.5th quantile) were aggregated and restricted to only a few portions of the genome. This was particularly striking for the most distant interspecific population pair (YP1 x MP6, Fig2A), where 314 of 404 (77%) 10kB windows with excess FST are located on LG2. Moreover, 290 of the outliers (71%) were in the peri-centromeric region between 41.985 and 43.175 cM on LG2. In contrast, all interspecific comparisons on the same side of the mountain pass exhibited excess divergence more dispersed across the genome (YP2 x MP4, Fig2B, FigS1 for all interspecific pair-wise comparisons). Of the 404 windows (10Kb) identified as outliers across all nine interspecific population pairs, only six (1.4%) were in common across all comparisons (Fig2A,B, FigS1). All six of these common outlier windows encompass the ROS or EL genes on LG2. Similarly, a contiguous block of ~210kB (8.9% of outliers) encompassing the *Am4’* CTG gene on scaffold316 and a ~190kB region of the adjacent scaffold829 (4.7%), were consistent outliers in seven of the nine comparisons. Only one other small 20kB region (0.004% of outliers) on LG1 (scaffold1187) was also consistently an outlier in seven of the nine comparisons. All remaining outliers (85%) were only in the upper 99.5th quantile in four or fewer interspecific comparisons.

*Drivers of genome-wide divergence*

Comparisons of nucleotide diversity *within* vs. absolute divergence *between* populations revealed that regions of excess FST are driven by shifts in both of these parameters (Fig 3). For the most distant population pair (YP1 x MP6; Fig 2A), the drivers of elevated FST (upper 99.5th quantile) depend on the genomic location. For example, around ROS, EL and FLS genes elevated FST is driven primarily by reduced polymorphism within one of the two populations (*w*) relative to the total diversity. In contrast, at the AM4’CGT gene this signal is not present, with excess FST driven by a combination of reduced *w* and increased *b* (Fig S2). Considering all regions of the genome with excess relative divergence (99.9th quantile) we see a deficit of high polymorphism *within* and *between* populations compared to regions of moderate FST (Fig. 3. See histogram compared with black line). This pattern is mostly driven by lower *w*in the *pseudomajus* populations relative to the *striatum* population on the other side of the mountain pass (YP1).

Among the regions of the genome with elevated FST driven instead by excess nucleotide divergence *b*), all of these occurred in linkage group 2, in close linkage with Am4’CGT. This coincides with a region of low recombination. We found a weak relation between lower recombination rate and excess FST but only on some chromosomes. The affect was strongest on LG2, with FST ranging from 0.01 to 0.96 for regions < 5 cM from centromere, compared with 0.01 to 0.36 for regions > 5 cM (Beta regression, Pseudo R2= 0.20, p < 2 x 10-12;Fig S3). However, on other chromosomes the signal was weak or absent (Fig S3).

*Strong allele frequency differences and clines strongly clustered across the genome*

Loci with strong allele frequency differences were detected on five of the eight chromosomes (*n* = 440 loci with > 0.9; Fig 4) or all eight when considering less differentiated loci (*n* = 1248 loci with > 0.8; Fig 4). In both cases, the majority were concentrated in just a few portions of the genome (see Fig 4). Some loci (6% at > 0.9, 25% at > 0.8) exhibited strong allele frequency differences aligned with the mountain pass, but lacked a monotonic decrease in allele frequency.

The remaining 414 highly divergent loci ( > 0.9) with allele frequencies characteristic of geographic clines, were highly associated with genes involved in flower pigmentation (Table 1). We found 208 loci (50% of 414) displayed clines were located less than 200kB from the structural gene involved in aurone pigmentation (*Am4’* CTG; scaffold316 on linkage group 2, 41.375 cM). A further 103 (25%) and 60 (15%) loci with steep clines were located on tightly linked to scaffold829 and scaffold60, respectively. The latter scaffold also contains the MUT gene known to affect anthocyanin pigmentation on the corolla lobes in *Antirrhinum*. Only 26 highly differentiated loci (6%) were detected near or within regulatory genes controlling anthocyanin intensity and patterning (ROS and EL) on LG6 (Fig 4), and nine (2%) in close proximity to FLS flavanol synthase on LG5).Lastly, steep clines were detected in close proximity to a gene involved in UDP-Glucose (on LG1), a native sugar donor for enzymes like Am4’CGT, UFGT, and 5GT in *Antirrhinum.* Other structural (e.g. CHS, CHI, F3H, DFR, ANS, AS1) and regulatory genes (DEL, Ve) involved in flower pigmentation in *Antirrhinum* exhibited no detectable steep clines in close mapping proximity (Fig 4) (Table 1). For each of these genes, the number of clines detected, declined with physical distance from coding regions (Fig S4). Only 44 (10%) of loci with steep clines were detected in locations unlinked to known flower colour genes (i.e. at least 1cM distant)

The distribution of widths across the genome did not significantly differ from a unimodal distribution (Fig 5, Hartigan’s test *D* = 0.0098, *P* = 0.9989). In contrast, cline centre across the genome displayed significant discordance among loci, departing significantly from a unimodal distribution (Fig 5, *D* = 0.0396, *P* = 2.2x10-12). Here, the majority of loci displayed a cline centre to the east of the phenotype cline, with a second smaller set of clines centred to the west. Only 29 (6.5%) of loci displayed clines within 1km of the phenotype cline, however 72% of these were linked to ROS or EL genes (Fig 5A, phenotype cline at ~12km). The discordance in cline centre among loci was strongly associated with genomic location, 298 of 300 loci (99%) with cline centre >13 km, occur in a small region at 43.18 cM on linkage group 2 near the *Am4’* CTG yellow gene (Fig S5). In contrast, loci with clines centre < 13 km, are distributed across the genome (Fig 5).

Re-genotyping the same individuals included in the pools reveals a high correlation in allele frequencies between pooled whole genomes and SNP genotypes (Fig Sxx). More intense spatial sampling (*n* = 1600 individuals across 83 demes) with SNP genotyping for a representative set of loci within each of these regions and genes, also confirmed the patterns observed in the pooled data, with characteristic monotonic decreases in allele frequency across the hybrid zone (Fig 7).

*Discordant cline centres at colour loci*

There was considerable staggering of cline centres at loci linked to genes related to magenta compared to yellow pigmentation. Cline centre for the regulatory anthocyanin genes (ROSmean = 11.8 km, ROSsd = 0.34; ELmean = 12.3 km, ELsd = 0.18) were located in approximately the same geographic area of the phenotypic cline. In contrast, loci at the structural aurone gene were staggered either side, depending on whether loci were upstream of the *Am4’* CTG yellow gene (Yellow\_upmean = 14.3 km, Yellow\_upsd = 0.49) or downstream (Yellow\_upmean = 9.1 km, Yellow\_upsd = 0.77). Across the genome, this abrupt fine-scale staggering of clines occurs only around the yellow structural gene *Am4’* CTG. This shift in cline centre occurs across the promoter, with all loci upstream of the *Am4’* CTG gene exhibiting cline centres > 12km whereas all loci downstream exhibit cline centres < 11 km (Fig 5,6). More intense spatial sampling with SNP genotyping confirmed the same discordance in cline centres among loci (Fig 7).

*Geographic and genomic clines detect flower colour loci as outliers*

Geographic clines (for loci with > 0.9) at the regulatory genes involved in anothocyanin pigmentation were substantially narrower (ROSmean = 3.2 km , ROSsd = 0.71; ELmean = 5.1 km, ELsd = 1.5) than the genomic average (mean = 6.8 km, sd = 1.5) (Fig 5A). Of the 24 loci in the upper 95% quantile of cline width, 10 (42%) were within the ROS or EL magenta genes and 9 (38%) were within ~2cM of the structural aurone *Am4’* CTG gene. While the remaining loci in the upper quantile had similarly narrow clines, all but one showed centres located more to the west (6-8 km), reflecting a sharp step in allele frequencies coinciding with the mountain pass (between YP1 and YP2, Fig 1). In contrast to ROS/EL loci, the majority of the loci most tightly linked to the yellow *Am4’* CTG gene (either immediately up or downstream) displayed cline widths that were indistinguishable from the genomic background (Fig 5A). Similar patterns were observed with more relaxed allele frequency differences, with the addition of several loci tightly linked to FLS and one locus linked to the gene involved in UDP-glucose synthesis, displaying narrower clines than the genomic background (Fig 5B).

Non-geographic (genomic clines) analyses showed similar results to the geographic approach. A total of six loci exhibited joint estimates of relative that were significantly different to the genomic background. Of these outlier loci, three were located within the ROS or EL gene on LG6 and displayed significantly steeper gradients (*u*) in allele frequencies with centre shifted (*v*) relative to the genome wide average. A further two outlier loci were tightly linked to the structural gene flavanol synthase (FLS) on LG5, and another loci linked to the aurone gene *Am4’* CTG exhibited similar gradients to the average but with significantly different centre (*v*) relative to the genome wide average (Fig 8).

*Associations between divergence and clines*

Regions of excess FST (99.5th quantile) corresponded to some of the locations where steep clines were detected. Using the most distant interspecific population pair (YP1 x MP6, Fig2A), 94 of 404 (23%) of 10kB windows with excess FST also possessed at least one locus with a steep cline. Of the 440 loci with steep clines, x % were within windows with excess FST, indicating that multiple loci with clines sit within some of the same 10kB windows. There was no clear relation between FST and geographic cline width, however the variability in widths diminished with increasing divergence (Fig S8).

Divergence and diversity statistics exhibited a number of differences among windows with and without steep clines. For comparisons with the outer most populations (YP1 x MP6), those 10 kB windows with clines displayed significantly higher FST, and substantially lower polymorphism within (*w*) the *striatum* population and *pseudomajus*, but no difference in divergence between (*b*) populations (Fig 9). For closer interspecific comparisons (e.g. YP2 x MP5) we found the same differences, in addition to a significantly higher divergence between (*b*) populations for windows harboring clines (FigS9).

*Discussion*

The distribution and properties of steep clines and genomic divergence in *Antirrhinum majus* is consistent with the difference between *A. m.* *pseudomajus* and *A. m.* *striatum* being primarily due to flower colour pattern differences*.* Using a novel method to efficiently estimate geographic cline parameters, we found steep clines to be restricted to very few portions of the genome, mostly within or in tight linkage with structural and regulatory genes that control patterns of anthocyanin and aurone pigmentation in *Antirrhinum*. Genomic divergence was similarly centred on these same regions, with the tail of the distribution enriched for regions containing these colour related genes. However, the pattern of diversity within and between populations suggests that these regions have a complex history, with different processes responsible for excess genomic differentiation across the genome. We focus our discussion on distinguishing between the drivers responsible for genome wide divergence and the patterns of clines across the genome. This enables a separation of the processes relating the origin of genomic divergence from those involved in generating contemporary barriers to gene flow.

*Islands of divergence*

Genomic divergence between these two subspecies of *Antirrhinum* is relatively low.

Measures of divergence between geographically isolated (FST = 0.057) and near adjacent *Antirrhinum* populations exchanging genes across the hybrid zone (FST = 0.037) are substantially lower than many classic systems in speciation genetics. For example, *Ficedula* flytcatchers , *Anopheles* mosquitos (FST = 0.21, Cruickshank & Hahn 2014) and some species in *Heliconius* (FST, Nadeau *et al.* 2012). The relative divergence in *Antirrhinum* is more similar to levels described among ecotypes such as *Timema* stick insects (FST = 0.03, Soria-Carrasco *et al.* 2014), *Astatotilapia* cichlids (FST = 0.038, Malinsky *et al.* 2015) and *Mimulus* monkeyflowers (FST= 0.081, Streisfeld *et al.* 2013) but greater than estimates among dune ecotypes in *Helianthus* sunflowers (FST = 0.004, Andrew *et al.* 2013). The genomic distribution of relative divergence varies considerably between these systems, from thousands of regions with elevated divergence across all chromosomes (e.g. *Timema*) to divergence concentrated in a few genomic locations associated with regions of low recombination (e.g. *Helianthus)*. Similar to the sunflowers, these *Antirrhinum* species exhibit highly heterogeneous patterns of divergence, with excess divergence concentrated in only a few portions of the genome.

Whether regions of excess divergence necessarily harbor genes affecting reproductive isolation has come increasingly under question (Noor & Bennett 2009; Turner & Hahn 2010; Cruickshank & Hahn 2014). Nevertheless, in *Antirrhinum* we found a strong correspondence between elevated FST and regions known to harbour genes responsible for phenotypic differences (e.g. *Rosea*, *Eluta*) and likely candidate genes given their known role in the flavonol biosynthetic pathway (Am4’CGT, FLS). The most consistent outlier FST windows (across all interspecific comparisons) mapped directly to the *Rosea* and *Eluta* genes. The second most consistent outlier regions map to the structural Am4’CGT gene involved in aurone (yellow) synthesis. Given these may be key genes that interact to determine the parental phenotypes (Schwinn *et al.* 2006; Ono *et al.* 2006; Whibley *et al.* 2006) and that these regions display steep clines in allele frequencies, FST scans work remarkably well in detecting loci that may be involved in phenotypic differences between these *Antirrhinum* species.

*Drivers of genomic divergence*

In *Antirrhinum*, the drivers of elevated FST depend on both the genomic region and the specific genes in which the sequences are embedded. For example, elevated FST at *Rosea* and the tightly linked *Eluta* gene are driven by reduced polymorphism within both parental populations. The surrounding sequences immediately outside these genes display average levels of polymorphism. This has been interpreted as a signal of multiple selective sweeps (Tavares et al., in prep). Similar reductions in polymorphism within populations are responsible for elevated FST at the Flavonol synthase gene (FLS). In contrast, the Am4’CGT gene displays no significant reduction in polymorphism within the gene compared to the surrounding sequence. Instead we observe overall lower polymorphism in the entire region and some signals of excess absolute divergence in the surrounding sequence. Arguably, this could be due solely to lower recombination, which could lead to fixed neutral differences, and hence the observed pattern even if Am4’CGT is not selected.

Considering the genome overall, excess relative divergence (99.9th quantile) was primarily due to reduced polymorphism within populations (Fig. 3. Histogram compared with black lines). This result is consistent with the majority of observations in other genome-wide studies (Cruickshank & Hahn 2014). In the case of *Antirrhinum*, the most extensive and heterogeneous levels of relative divergence involved comparisons with the *striatum* population from over the mountain pass (YP1, Fig1, Fig3). Here, the overall higher levels of nucleotide diversity (*w*) at YP1, drives excess divergence among *pseudomajus* and other *striatum* populations. Although absolute sequence divergence (*b*) is responsible for excess FST in some regions of the genome, all of these occurred in linkage group 2, in tight linkage with Am4’CGT. Further understanding of these patterns will require more widespread geographic sampling of *pseudomajus* and *striatum* populations, and utilizing methods to distinguish between the role of demography, drift and selection between closely related species (e.g. Joint site frequency spectrum; Chen *et al.* 2007).

*Clines associated with flower colour genes*

We found evidence of selection and barriers to gene flow acting predominately at flower genes. Theory predicts that at equilibrium, diagnostic loci that are under strong selection will display steeper clines than the neutral background. We detected steep clines at both regulatory and structural genes involved in the flavonol biosynthetic pathway, with few diagnostic loci displaying clines elsewhere in the genome. Importantly, using a new method for fitting clines we detected the same steep clines previously identified at *Rosea* and at the tightly linked *Eluta* gene. In addition to *Rosea* and *Eluta* (LG6), steep clines were present at other structural genes involved aurone (Am4’CGT on LG2) pigmentation and the broader flavonol pathway (FLS on LG5, UDP-Glucose on LG1) in *Antirrhinum*. While we do not have direct evidence for phenotypic effect of these genes unlike *Rosea* and *Eluta*, they play key roles in the flavonol pathway where divergent haplotypes may harbor substitutions that generate phenotypic differences. Given that the *Antirrhinum* hybrid zone has likely been present for over 100 generations (Tavares et al., in prep) these patterns are unlikely to represent non-equilibrium conditions. Although steep clines are in regions enriched for flower colour genes, only a fraction of the possible structural or regulatory genes identified in *Antirrhinum* exhibit steep clines. This confirms previous suggestions that relatively few genes may be involved in transitions from magenta to yellow pigmentation in *Antirrhinum* (Schwinn *et al.* 2006; Whibley *et al.* 2006). Although these subspecies predominately differ only in flower colour, steep clines in genomic regions unrelated to colour provides scope for future association studies with other traits that may contribute to reproductive isolation.

Differences in the geographic position of clines at independent isolating genes also provide important insight into the role of gene interactions. With both anthocyanin and aurone pigmentation interacting to constitute the parental forms, theory predicts these loci will be synergistically coupled (concordant cline centres) upon secondary contact due to the influx of parental gene combinations and linkage disequilibrium (Barton & Hewitt 1985). Alternatively, if recombinants have different fitnesses, due to epistasis between loci (as in Dobzhansky-Muller models), then clines will be staggered (Gavrilets 1997) with the fitter recombinants tending towards the centre of the hybrid zone (e.g. Searle 1986; Virdee & Hewitt 1994). Our current results support the latter, with evidence of staggered clines among loci involved in anthocyanin and aurone pigmentation. Although the staggered clines could be an artifact of the coarse spatial sampling and the cline approximation method, current efforts using spatially dense sampling and SNP genotyping indicate that these patterns are real features of this hybrid zone (D Field unpublished data). Whether these patterns reflects a consistent mechanism of reproductive isolation or the scattering of clines due to drift will require further investigations of cline positions in replicate hybrid zones.

Our data highlights that the recombination landscape in which selected loci are embedded is important in determining the distribution of clines across the genome. We found higher frequencies of steep clines clustered around the structural aurone related gene (*Am4’* CTG) compared with regulatory and structural genes controlling anthocyanin pigmentation (*Rosea*, *Eluta*, FLS). Given that local recombination around *Am4’* CTG is dramatically lower than either of these other regions, (*Am4’* CTG ~7Mb/1cM, *Rosea*/*Eluta* ~300Kb/1cM), the simplest explanation is enhanced hitchhiking of neutral loci. Indeed, the difference in the number of steep clines around these two regions is roughly proportional to the difference in recombination rates (i.e. 11% fewer loci with clines at ROS/EL compared to *Am4’* CTG). The relation with cline density and recombination rate is similar to the observation that elevated FST is often associated with regions of the genome with lower rates of recombination (McGaugh *et al.* 2012; Cruickshank & Hahn 2014). For clines, lower recombination rates increase the target size of selection and also slow the exchange of alleles between neutral and linked variants across the physical genome. Consequently, selection or neutrality may contribute to this pattern.

*Correspondence between divergence and steep clines*

Divergent selection in the presence of gene flow or selective sweeps in allopatry can result in exceptional levels of genomic divergence. If these genomic regions also contribute to reproductive isolation, we expect that secondary contact in hybrid zones will also generate steep clines in similar regions of the genome. In the *Antirrhinum* hybrid zone, we found a general correspondence between regions of excess FST and the presence of steep clines. This was especially the case around the *Rosea* and *Eluta* genes, which were the only regions that exhibited steep clines and excess divergence (upper 99.95th quantile) in all pair-wise population comparisons. A similar pattern was also present at the Am4’CGT gene involved in aurone pigmentation, where the majority of clines across the genome were located. However, many regions of excess FST exhibited no steep clines (75%). Although it may be tempting to attribute the lack of steep clines to signal a genomic region unrelated to reproductive isolation, clines can only be fitted at loci with strong allele frequency differences (). Consequently, it could also be that divergence is due to smaller allele frequency differences that reflect polygenic adaptation that could still contribute to reproductive isolation yet display no steep clines.

If divergence were neutral, we may expect clines and divergence to become decoupled following secondary contact. In some systems, there is a coarse correspondence between genomic islands of FST and QTLs for species defining traits (e.g. Rogers & Bernatchez 2007; Via 2012) and reproductive isolation (e.g. Payseur *et al.* 2004). Only a few studies have directly compared FST to measures of introgression in admixed populations for a relatively large number of loci (i.e. 1000s) (Gompert *et al.* 2012; Nosil *et al.* 2012; Parchman *et al.* 2013). Using genomic cline approaches, these studies report overall weak to moderate correlations between FST and gradient parameters (method of Gompert & Buerkle 2010) that were, in some cases, inconsistent across hybrid zones. However, one limitation of these studies is that they lacked knowledge on the genomic location of QTLs for reproductive isolation. Our results are generally consistent with these studies, with correspondence in some genomic regions but not in others. However, if we consider only the most consistent outlier FST regions among multiple interspecific pairs, these all correspond to major flower colour genes. The strong association between divergence, clines and flower colour genes in *Antirrhinum* may reflect the relatively simple genetic differences between *A. m. pseudomajus* and *A. m.* *striatum*.

*Utility of scanning genomes for steep clines to locate barrier genes*

The approach of scanning genomes for steep clines and mapping genes responsible for reproductive isolation is becoming increasingly feasible with the availability of whole genome data. In the *Antirrhinum* hybrid zone, where the link from genotype to phenotypes under selection is well understood, we found that the majority of clines are located in tight linkage with flower colour genes involved in anthocyanin or aurone pigmentation. Moreover, when considering only highly divergent loci (), some flower colour genes (*Rosea* and *Eluta*) are among the steepest gradients in allele frequencies as a function of distance (*geographic* *clines*) and emerge as statistical outliers in relation to the hybrid index (*genomic* *clines*) compared to the genomic background. However, with the inclusion of less divergent loci (), it becomes increasingly difficult to identify these loci from the genomic background. This is due, in part, to one of the limitations of this cline approximation method, where fluctuations in allele frequencies and cline reversals can substantially narrow cline widths. However, by sampling larger number of populations across a hybrid zone with a subset of loci we have verified that the general pattern of clines remains across the genome.

We conclude by identifying several problems for future research in understanding genome wide distributions of clines. Firstly, the covariance of cline parameters due to tight linkage in *Antirrhinum* highlights the need for future theoretical approaches that deal with the issue of non-independence among loci. Unlike classical hybrid zone studies using small numbers of markers, the move to whole genomes necessarily results in loci in tighter linkage and thus many of them will not be statistically independent. This is also an issue for genomic cline methods, which are based on detecting outliers from the neutral background loci, which are wrongly presumed independent. One approach to dealing with this problem is to move beyond single loci and follow blocks of genome through space (Sedghifar *et al.* 2015). Secondly, we still lack a theoretical understanding of the expected distribution of clines across genome wide data and how the rate of false positives depends on the time since secondary contact, drift and the strength of selection. Lastly, determining how far genetic barriers persist at neutral sites around a selected locus will also provide insight into the processes structuring genome wide divergence. This will enable the further dissection of the role of stochastic, historical and contemporary forces in driving patterns of divergence and clines across genomes.

**Acknowledgements**

**References**

Andalo C, Cruzan MB, Cazettes C *et al.* (2010) Post-pollination barriers do not explain the persistence of two distinct *Antirrhinum* subspecies with parapatric distribution. *Plant Systematics and Evolution*, **286**, 223–234.

Andrew RL, Kane NC, Baute GJ, Grassa CJ, Rieseberg LH (2013) Recent nonhybrid origin of sunflower ecotypes in a novel habitat. *Molecular Ecology*, **22**, 799–813.

Barton N (1986) The effects of linkage and density-dependent regulation on gene flow. *Heredity*, **57**, 415–426.

Barton NH, Gale KS (1993) Genetic analysis of hybrid zones. In: *Hybrid zones and the evolutionary process (ed. R.G. Harrison)*, pp. 13–45. Oxford University Press, Oxford, UK.

Barton NH, Hewitt GM (1985) Analysis of hybrid zones. *Annual Review of Ecology and Systematics*, **16**, 113–148.

Bradshaw HD, Schemske DW (2003) Allele substitution at a flower colour locus produces a pollinator shift in monkeyflowers. *Nature*, **426**, 176–178.

Charlesworth B (1998) Measures of divergence between populations and the effect of forces that reduce variability. *Molecular Biology and Evolution*, **15**, 538–543.

Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics*, **134**, 1289–1303.

Chen H, Green RE, Pääbo S, Slatkin M (2007) The Joint Allele-Frequency Spectrum in Closely Related Species. *Genetics*, **177**, 387–398.

Cribari-Neto F, Zeileis A (2010) Beta Regression in R. *Journal of Statistical Software*, 1–24.

Cruickshank TE, Hahn MW (2014) Reanalysis suggests that genomic islands of speciation are due to reduced diversity, not reduced gene flow. *Molecular Ecology*, **23**, 3133–3157.

Ellis TJ, Field DL (in prep) Pollinator behaviour causes assortative mating and frequency-dependent pollinator selection for flower colour in a snapdragon (Antirrhinum) hybrid zone.

Endler J. (1977) *Geographic variation, speciation, and clines*. Princeton University Press.

Fitzpatrick BM (2013) Alternative forms for genomic clines. *Ecology and Evolution*, **3**, 1951–1966.

Futschik A, Schlötterer C (2010) The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, **186**, 207–218.

Gavrilets S (1997) Hybrid zones with Dobzhansky-type epistatic selection. *Evolution*, 1027–1035.

Gay L, Crochet P-A, Bell DA, Lenormand T (2008) Comparing Clines on Molecular and Phenotypic Traits in Hybrid Zones: A Window on Tension Zone Models. *Evolution*, **62**, 2789–2806.

Gompert Z, Buerkle CA (2010) introgress: a software package for mapping components of isolation in hybrids. *Molecular Ecology Resources*, **10**, 378–384.

Gompert Z, Lucas LK, Nice CC *et al.* (2012) Genomic regions with a history of divergent selection affect fitness of hybrids between two butterfly species. *Evolution*, **66**, 2167–2181.

Haldane JBS (1948) The theory of a cline. *Journal of Genetics*, **48**, 277–284.

Lynch M, Bost D, Wilson S, Maruki T, Harrison S (2014) Population-Genetic Inference from Pooled-Sequencing Data. *Genome Biology and Evolution*, **6**, 1210–1218.

Malinsky M, Challis RJ, Tyers AM *et al.* (2015) Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science*, **350**, 1493–1498.

McGaugh SE, Heil CSS, Manzano-Winkler B *et al.* (2012) Recombination Modulates How Selection Affects Linked Sites in Drosophila. *PLOS Biol*, **10**, e1001422.

Nadeau NJ, Whibley A, Jones RT *et al.* (2012) Genomic islands of divergence in hybridizing Heliconius butterflies identified by large-scale targeted sequencing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 343–353.

Noor M a. F, Bennett SM (2009) Islands of speciation or mirages in the desert? Examining the role of restricted recombination in maintaining species. *Heredity*, **103**, 439–444.

Nosil P, Parchman TL, Feder JL, Gompert Z (2012) Do highly divergent loci reside in genomic regions affecting reproductive isolation? A test using next-generation sequence data in Timema stick insects. *BMC Evolutionary Biology*, **12**, 164.

Ono E, Fukuchi-Mizutani M, Nakamura N *et al.* (2006) Yellow flowers generated by expression of the aurone biosynthetic pathway. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 11075–11080.

Parchman TL, Gompert Z, Braun MJ *et al.* (2013) The genomic consequences of adaptive divergence and reproductive isolation between species of manakins. *Molecular Ecology*, **22**, 3304–3317.

Payseur BA, Krenz JG, Nachman MW (2004) Differential patterns of introgression across the X chromosome in a hybrid zone between two species of house mice. *Evolution*, **58**, 2064–2078.

Polechová J, Barton N (2011) Genetic drift widens the expected cline but narrows the expected cline width. *Genetics*, **189**, 227–235.

Rausher MD (2008) Evolutionary Transitions in Floral Color. *International Journal of Plant Sciences*, **169**, 7–21.

Rogers SM, Bernatchez L (2007) The Genetic Architecture of Ecological Speciation and the Association with Signatures of Selection in Natural Lake Whitefish (Coregonus sp. Salmonidae) Species Pairs. *Molecular Biology and Evolution*, **24**, 1423–1438.

Schwinn K, Venail J, Shang Y *et al.* (2006) A Small Family of MYB-Regulatory Genes Controls Floral Pigmentation Intensity and Patterning in the Genus Antirrhinum. *The Plant Cell Online*, **18**, 831–851.

Searle JB (1986) Factors Responsible for a Karyotypic Polymorphism in the Common Shrew, *Sorex araneus*. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **229**, 277–298.

Sedghifar A, Brandvain Y, Ralph P, Coop G (2015) The Spatial Mixing of Genomes in Secondary Contact Zones. *Genetics*, **201**, 243–261.

Slatkin M (1973) Gene Flow and Selection in a Cline. *Genetics*, **75**, 733–756.

Soria-Carrasco V, Gompert Z, Comeault AA *et al.* (2014) Stick Insect Genomes Reveal Natural Selection’s Role in Parallel Speciation. *Science*, **344**, 738–742.

Streisfeld MA, Young WN, Sobel JM (2013) Divergent Selection Drives Genetic Differentiation in an R2R3-MYB Transcription Factor That Contributes to Incipient Speciation in *Mimulus aurantiacus*. *PLoS Genet*, **9**, e1003385.

Szymura JM, Barton NH (1986) Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegata*, near Cracow in Southern Poland. *Evolution*, **40**, 1141–1159.

Tavares H, Field DL, Whibley A *et al.* (in prep) Adjacent genomic islands identify selection on Antirrhinum flower colour patterns.

Turner TL, Hahn MW (2010) Genomic islands of speciation or genomic islands and speciation? *Molecular Ecology*, **19**, 848–850.

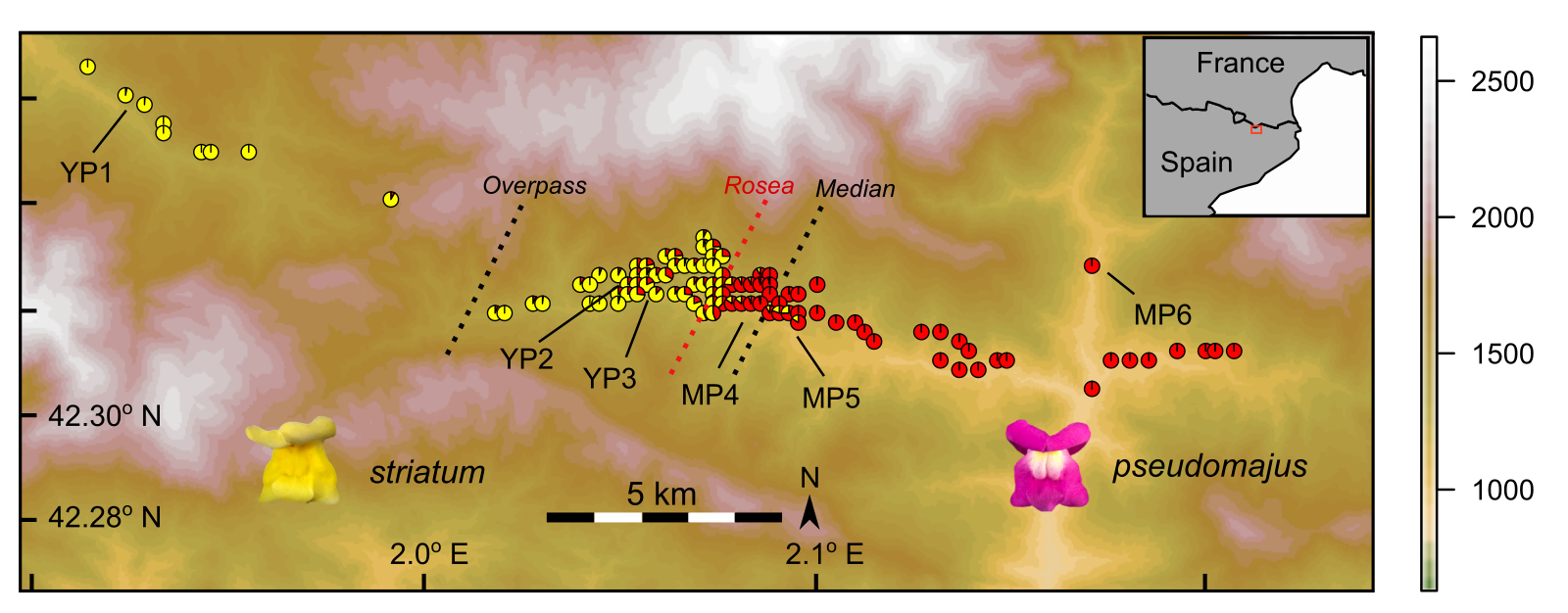
Turner TL, Hahn MW, Nuzhdin SV (2005) Genomic Islands of Speciation in Anopheles gambiae. *PLoS Biol*, **3**, e285.

Via S (2012) Divergence hitchhiking and the spread of genomic isolation during ecological speciation-with-gene-flow. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **367**, 451–460.

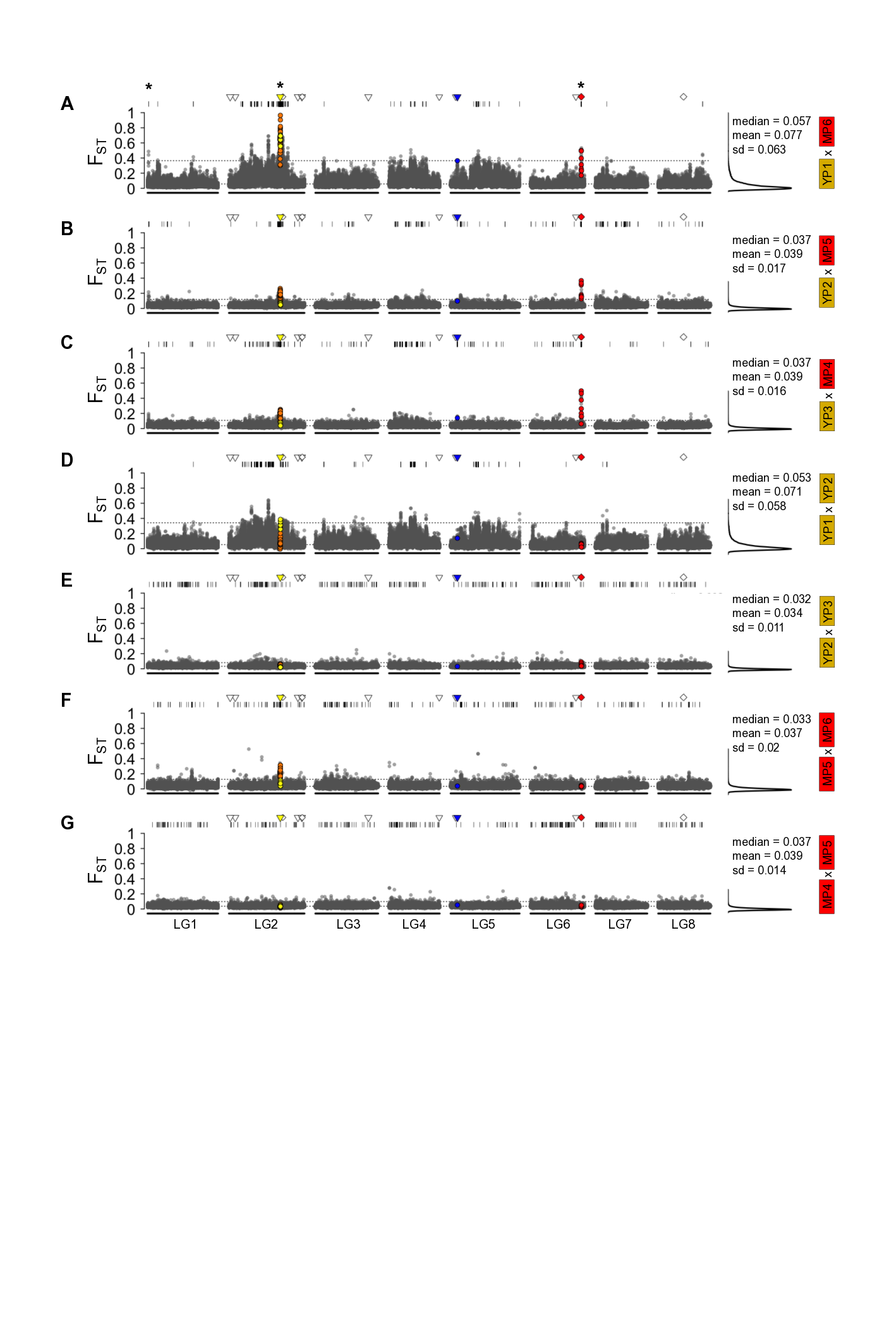
Virdee SR, Hewitt GM (1994) Clines for Hybrid Dysfunction in a Grasshopper Hybrid Zone. *Evolution*, **48**, 392–407.

Whibley AC, Langlade NB, Andalo C *et al.* (2006) Evolutionary Paths Underlying Flower Color Variation in Antirrhinum. *Science*, **313**, 963–966.

Wu C-I (2001) The genic view of the process of speciation. *Journal of Evolutionary Biology*, **14**, 851–865.

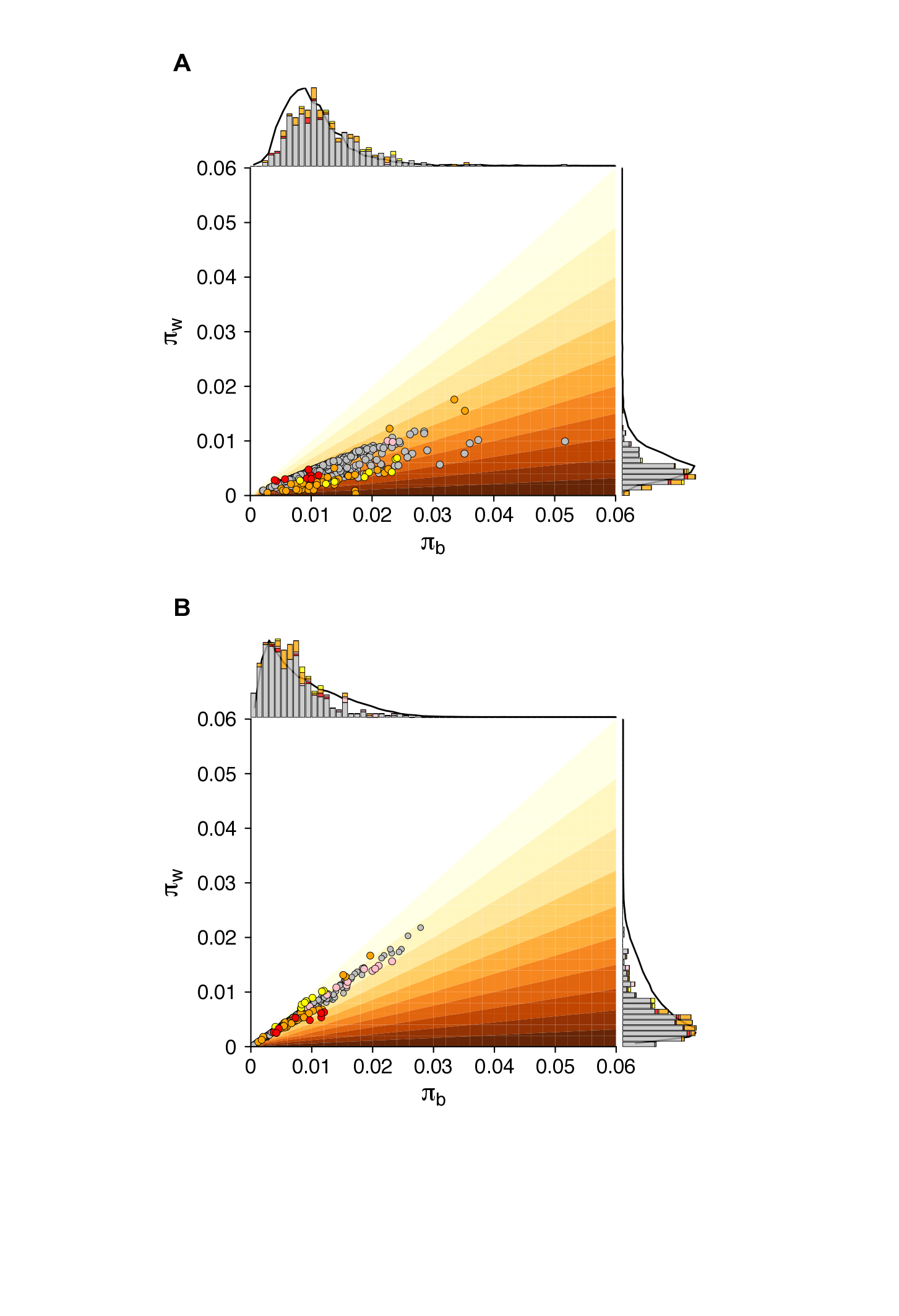


**Figure 1.** Sampling locations for individuals and pooled genomes across the *Antirrhinum majus* hybrid zone that occurs from west (*A. m.* *striatum*) to east (*A. m.* *pseudomajus*) in the Spanish Pyrenees. Each pie diagram shows the proportion of *ROS* (*pseudomajus -* red) compared to *ros* (*striatum* - yellow) alleles at the *Rosea* locus within a 200 metre diameter (see methods). Labels refer to the location and names of the six pools of 50 individuals used for whole genome sequencing. Dashed lines indicate the approximate centre of the *Rosea* cline (red), the genome-wide median cline centre (black - median) and the remaining divergent loci centred on the mountain pass (black – overpass).

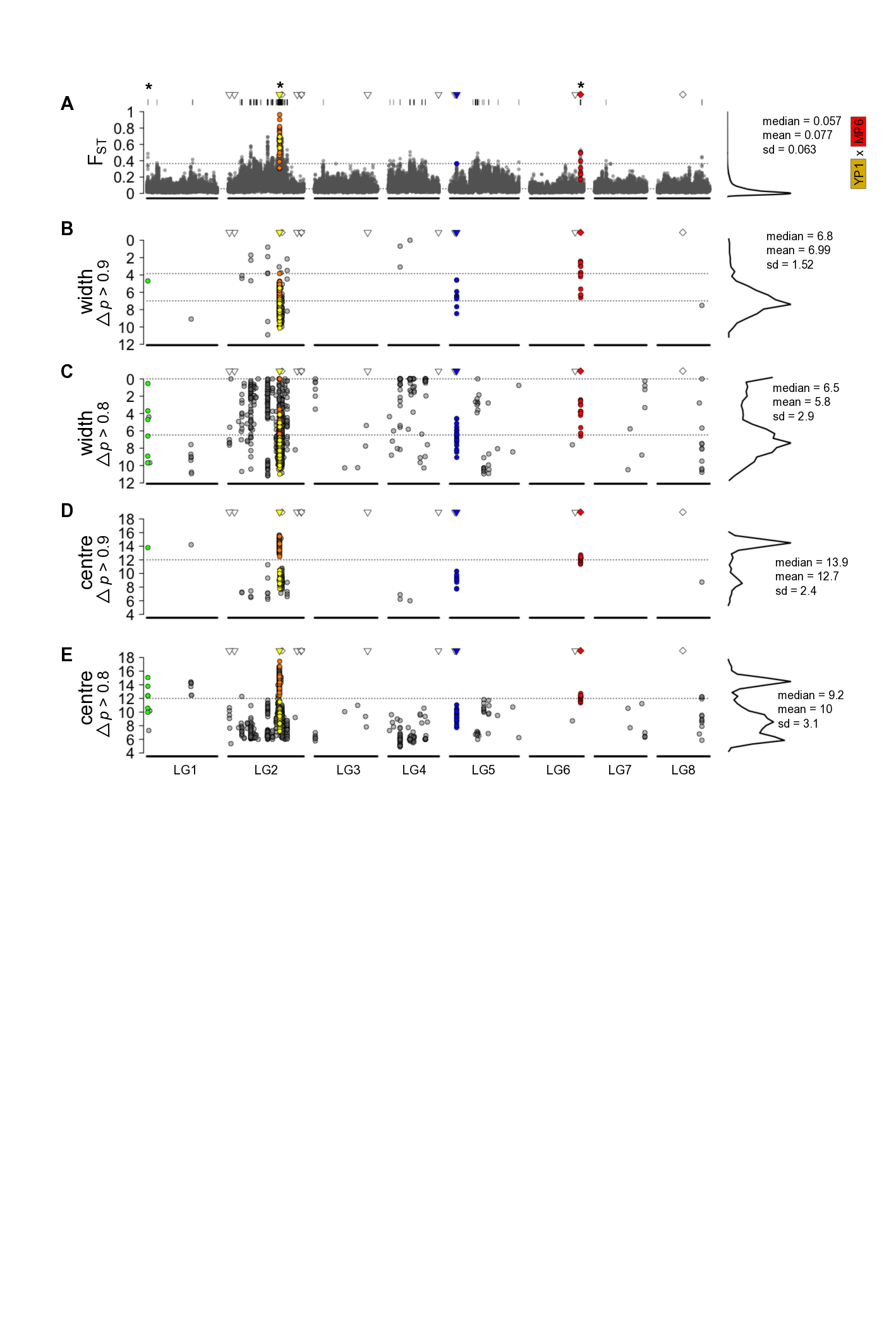


**Figure 2**. Relative divergence (FST) for 10kB windows along the eight linkage groups of *Antirrhinum majus*. The FST panels include three inter-subspecific population pairs (*A. m. striatum* x *A. m. pseudomajus* pairs (A-C), two *striatum* pairs (D,E) and two *pseudomajus* pairs (F,G). For each FST plot, windows are classified as background (grey), tightly linked (<300kB) to Am4’CTG (yellow = downstream/orange = upstream), flavonol synthase FLS (blue), UDP-Glucose (green) or the ROS/EL genes (red). Locations of genes involved in flower pigmentation in *Antirrhinum* indicated above each panel as structural (triangle) or regulatory (diamond) with the same colours (as above) if clines present and white if absent. Windows with excess FST (above 99.5th quantile) indicated with black lines above each panel. The median and 99.5th quantile also indicated within each plot (dashed lines). Asterisk above panel A indicates genomic regions that displayed consistent excess FST across multiple inter-subspecific pairwise comparisons. Positions of windows in physical distance (bp) along linkage groups.

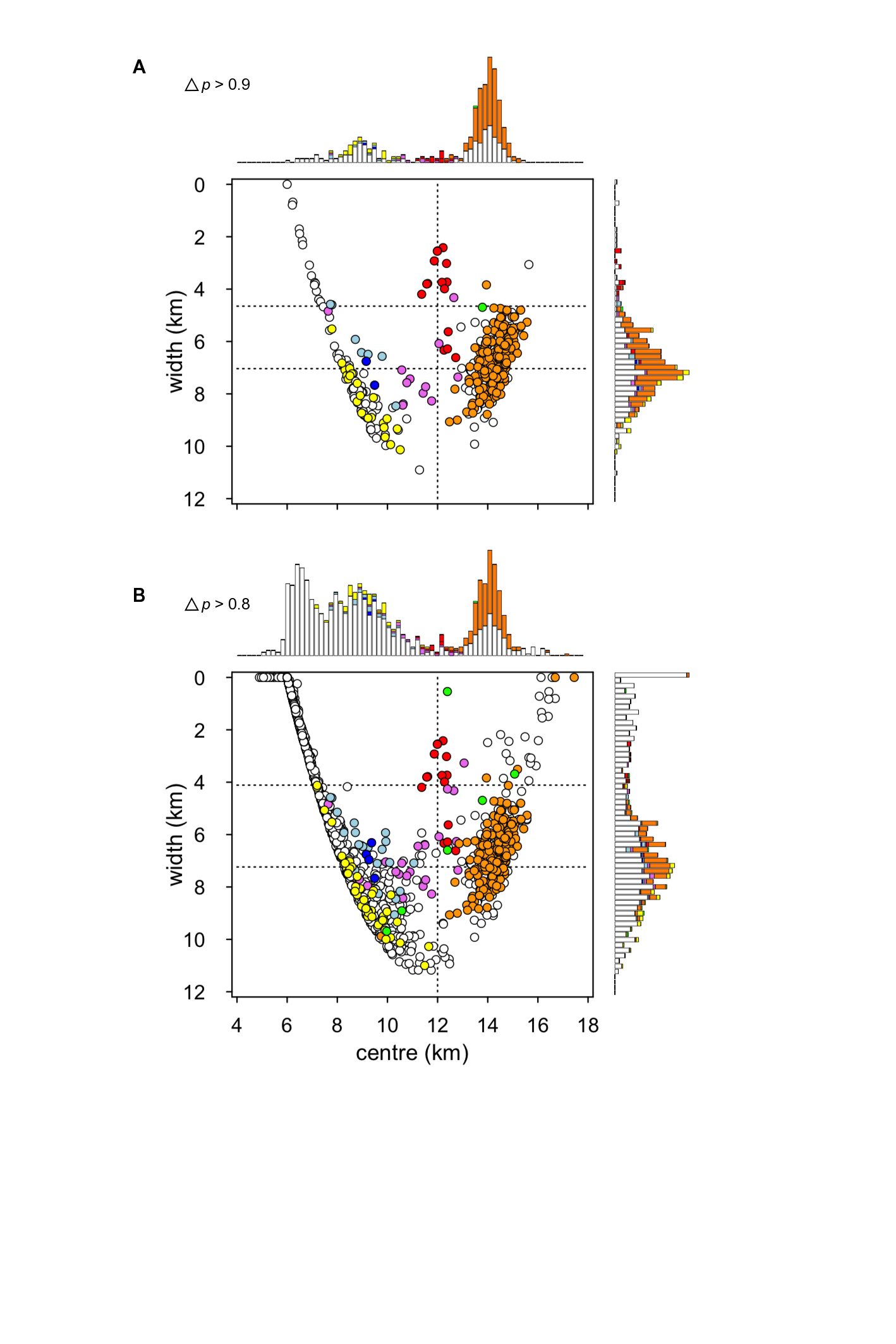
[move Supplement]



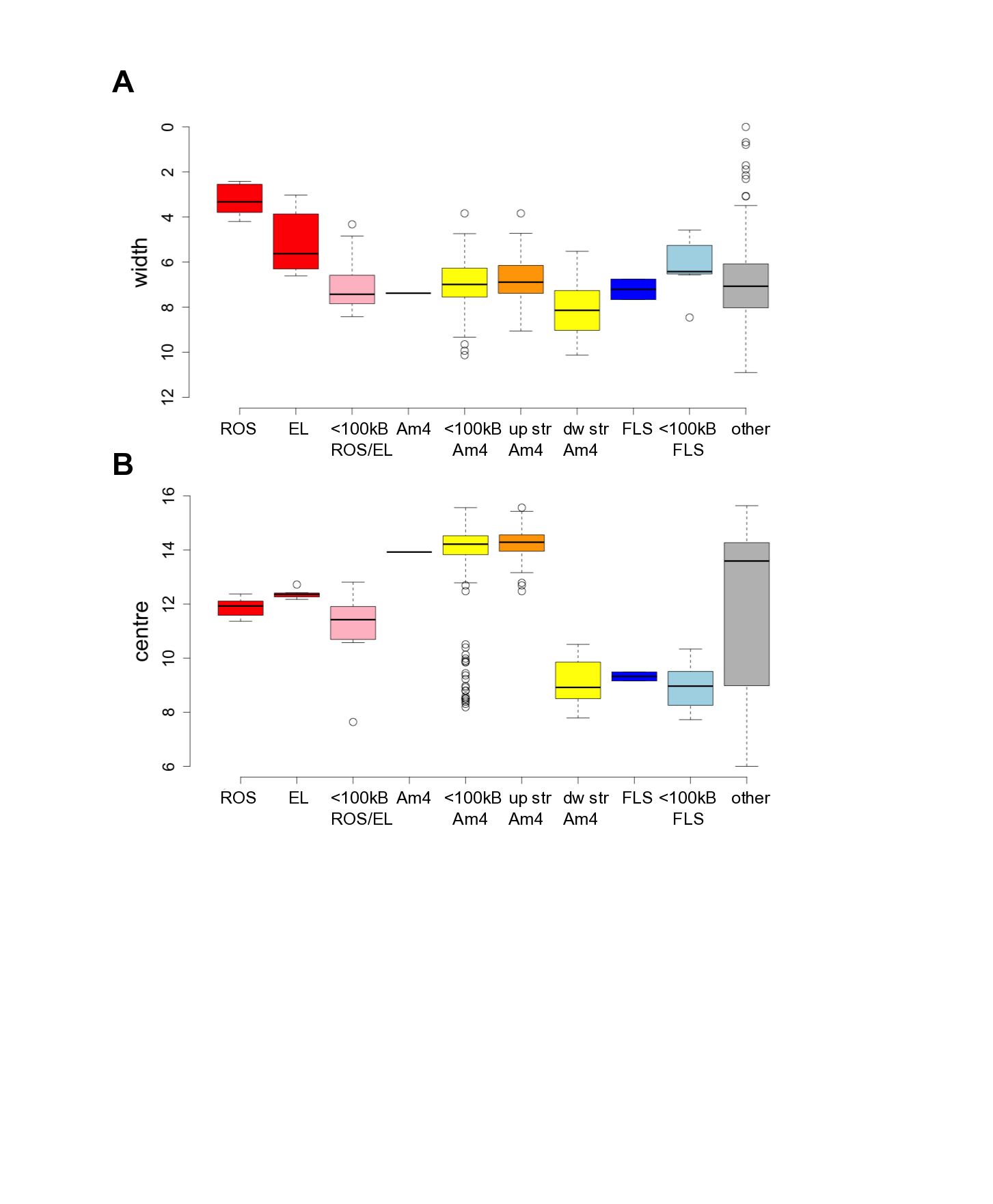
**Fig 3.** Diversity *within* (*w*) vs. diversity *between* (*b*) for pooled whole genome data for *pseudomajus* x *striatum* pairs YP1 x MP6 (A) and YP2 x MP5 (B) for outlier FST 10Kb windows across the genome. FST contours shown in 0.1 increments, with darker colours indicating higher values. Red circles = windows that overlap with ROSEA or ELUTA, pink = region between ROSEA and ELUTA, yellow = downstream of Am4’CGT, orange = upstream of Am4’CGT. Histograms indicate the frequency distribution of the outliers (99.95th quantile in FST), and the black lines behind the histograms indicate the frequency of moderate FST (median to upper 99.95th quantile).

****

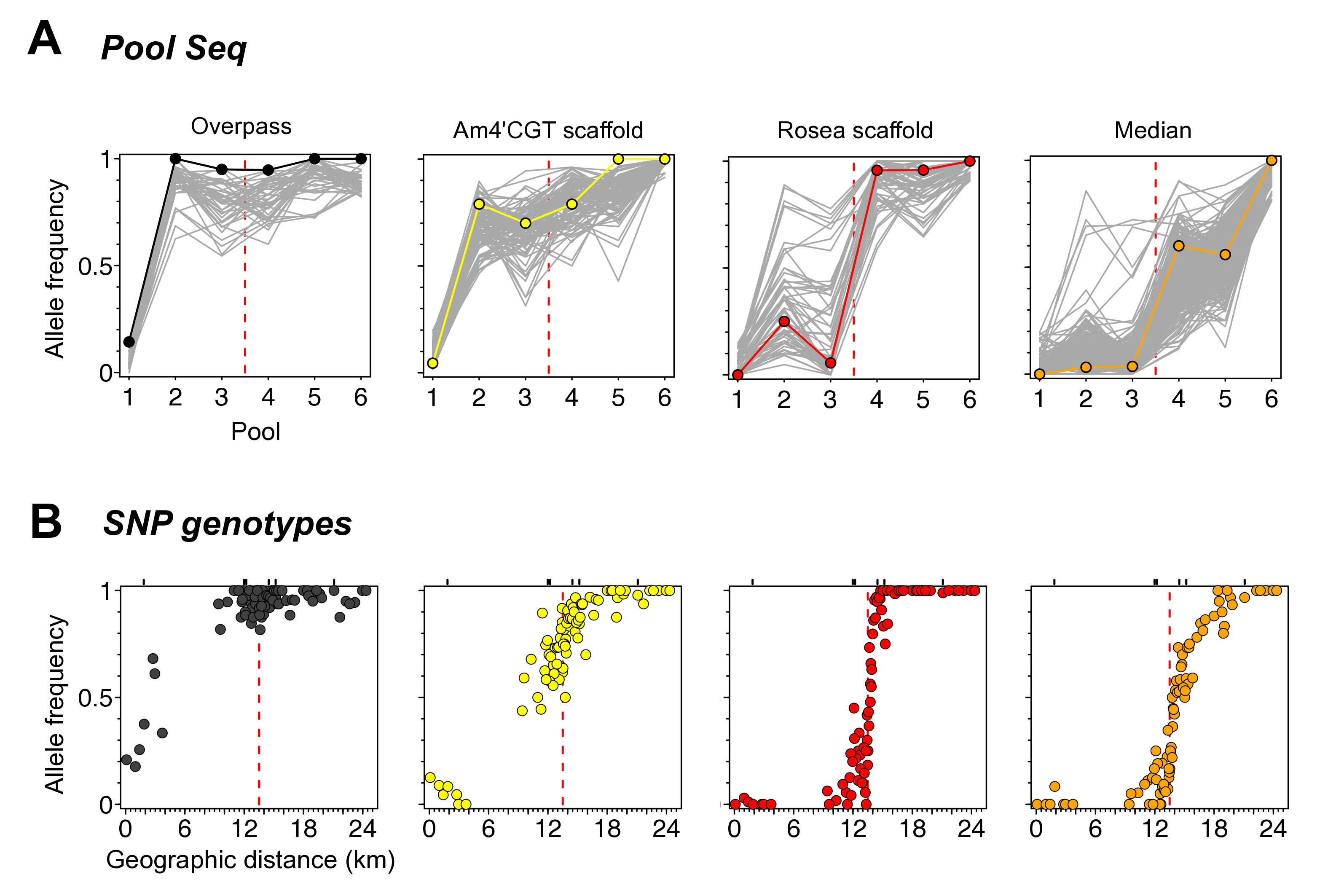
**Fig 4**. Relative divergence (FST) for 10kB windows (A) and geographic cline width (B, C) and centre (D, E) for individual loci along the eight linkage groups of *Antirrhinum majus*. The FST panel is for the most distant inter-subspecific population pair (*A. m. striatum* x *A. m. pseudomajus*). Windows are classified as background (grey), tightly linked (<300kB) to Am4’ CTG (yellow = downstream/orange = upstream), flavanol synthase FLS (blue), gene involved in UDP-Glucose (green) or the ROS/EL genes (red). Locations of flower colour genes indicated above each panel as structural (triangle) or regulatory (diamond) with the same colours (as above) if clines present and white if absent. Windows with excess FST (above 99.5th quantile) indicated with black lines above each panel. The median and 99.5th quantile also indicated within each plot (dashed lines). Cline estimates (in kilometres) estimated for all loci with allele frequency differences in the outer pool, ≥ 0.9 (B, D) and ≥ 0.8 (C, E). For cline width, 95% quantiles and median indicated (dashed lines), for cline centre position of phenotype cline indicated (dashed line). Asterisk above panel A, indicates genomic regions that were consistent outliers across multiple interspecific pairwise comparisons. Positions of windows and clines in physical distance (bp) along linkage groups.



**Figure 5**. Joint genome-wide point estimates of geographic cline centre and width (in kilometres) at the *Antirrhinum* hybrid zone for loci with allele frequency differences among the outer pools of (A)>0.9 (*n* = 415 loci) and (B)>0.8 (*n* = 930 loci). For each, central figure, red = loci within ROS/EL complex on LG6, pink = loci <300Kb from ROS/EL, yellow = loci downstream of *Am4’* CTG on LG2 (yellow determining gene), orange = loci upstream of *Am4’* CTG (yellow determining gene), dark blue = FLS gene on LG5, light blue = loci <300Kb from FLS, green = gene involved in UDP-Glucose, white = loci unlinked to known structural or regulatory genes involved is flower pigmentation. Horizontal dashed lines indicate median and upper 95% quantile for cline widths, vertical line approximate location of phenotypic cline in anthocyanin intensity. Histograms on the right and above of central figure indicate the frequencies of loci at 200 intervals in width and centre (the same colours as described above).



**Figure 6**. Comparison of cline widths (A) and centres (B) for loci within known regulatory or structural flower colour genes (ROS/EL, Am4= Am4’CTG, FLS), loci that are tightly linked to these genes (<100kB ROS/EL, Am4, FLS), upstream or downstream of Am4CTG (up str Am4, dw str Am4) or other regions > 500kB from known colour related genes.



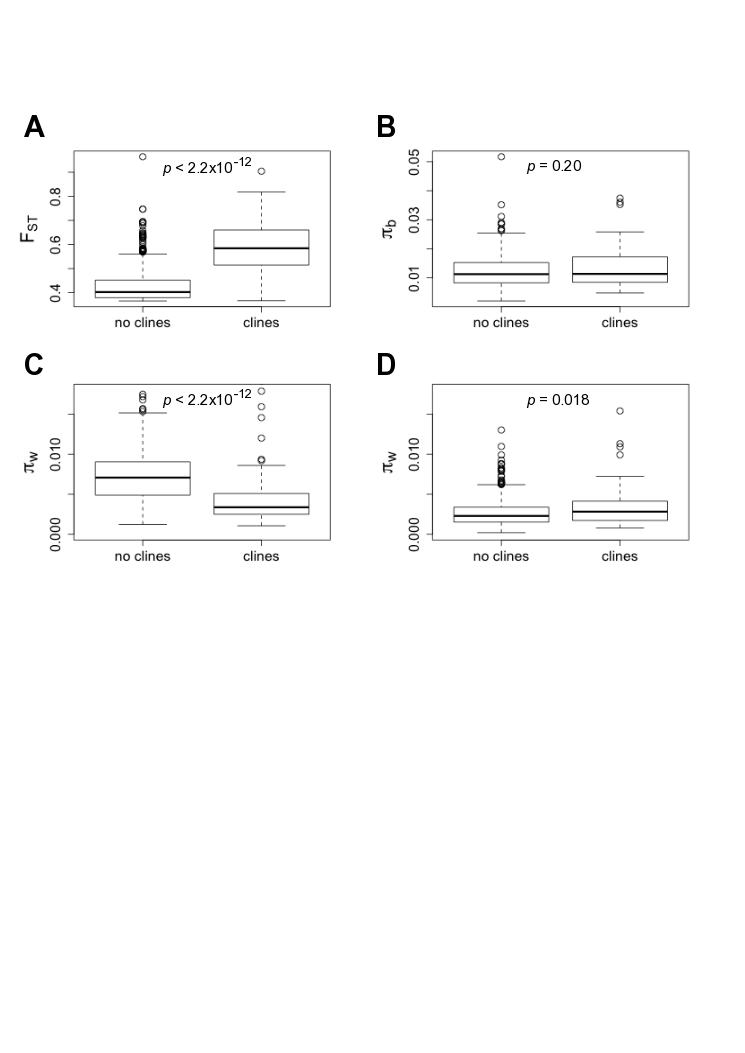
**Figure 7**. Allele frequencies for (A) whole-genome Pool Seq data and (B) corresponding SNP genotypes for a subset of loci for *n* = 1600 individuals. Four representative genomic regions include; those where allele frequencies transition across (i) the mountain pass (Overpass), (ii) to the west of the magenta phenotype cline (Am4’CGT scaffold), (iii) in line with magenta phenotype cline (Rosea scaffold) and (iv) where the bulk of allele frequencies transition (Median). One example locus was genotyped in each of these genomic regions, and the corresponding allele frequencies in the pool seq is indicated in colour compared to the background loci on the same scaffold (grey lines A). Location of the six pools relative to intensive sampled individuals indicated with black lines above (B).



**Figure 8**. Logit-logistic (non-geographic) genomic clines at the hybrid zone for the most divergent loci (>0.9) across the *Antirrhinum* genome. Grey lines are the genomic background, and coloured lines indicate the statistical outliers and refer to loci tightly linked to the ROS/EL (red), *AM4’* CTG (orange) and flavanol pathway gene FLS (black). The parameter *u* is the relative difference in cline position and *v* gives the relative difference in cline gradient compared to the genome wide average.

**Note: current version for all loci. I will generate a second panel (column) with picking only one locus per 0.5cM in an attempt to account for the issue of tight linkage and representation of the genomic average.**

Move to supplement or merge 7 and 8



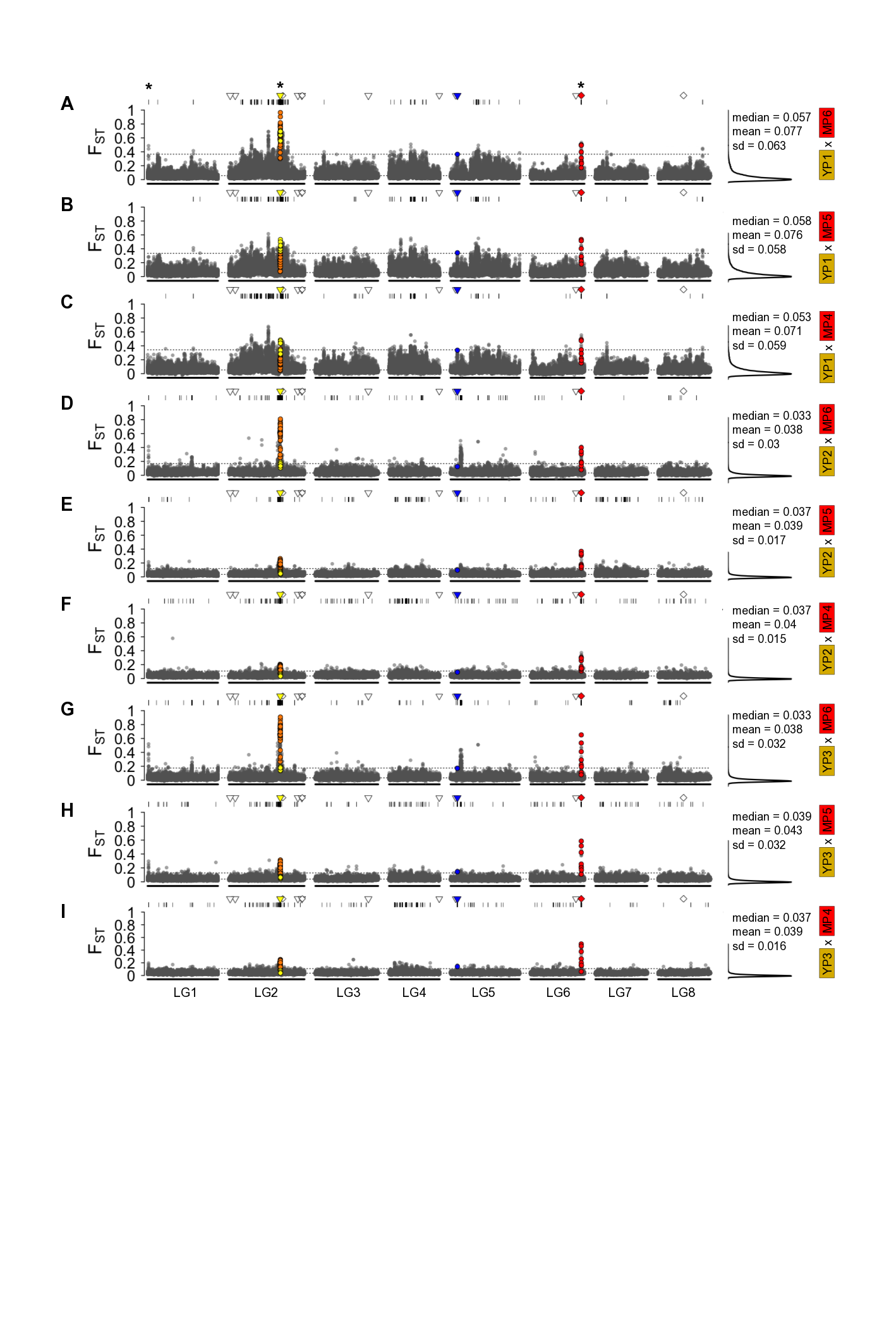
**Fig 9.** Comparative tests of 10kB windows harbouring steep clines vs. those without clines for the most distant pairwise comparison (YP1 x MP6). Measures of relative divergence FST (A), average polymorphism between populations *b* (B), and polymorphism within *w* for striatum (C) and pseudomajus (D). Tests for significant differences from non-parametric Mann-Whitney U test.

Also do this for closer pools

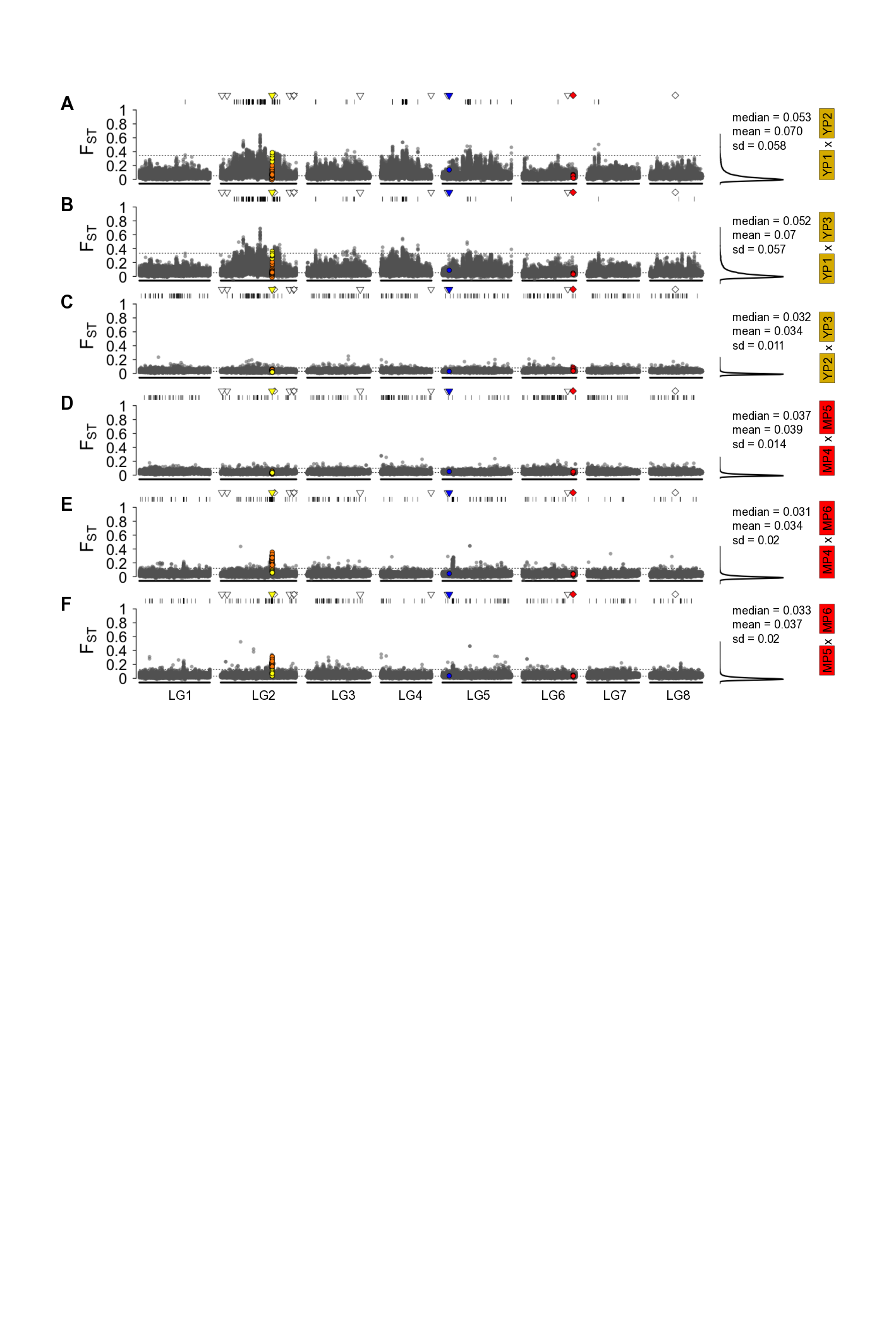
**Table 1. Flower colour genes**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Gene | Colour | Type | LG | cM | scaffold | # clines  ∆p16≥0.9  *Coding* | # clines  ∆p16≥0.9  *<200kB* | # clines  ∆p16≥0.8  *Coding* | # clines  ∆p16≥0.8  *<200kB* | FST  *Coding* | FST  *<200kB* | πb  *Coding* | πb  *<200kB* | πw  *Coding* | πw  *<200kB* |
| MUT | M | R | 2 | 43.18 | scaffold60 | 26 (0) | 13 (0) |  |  |  |  |  |  |  |  |
| DEL | M | R | 2 | 75.07 | scaffold591 |  |  |  |  |  |  |  |  |  |  |
| MIXTA | M | R | 3 | 32.67 | scaffold699 |  |  |  |  |  |  |  |  |  |  |
| PAL | M | R | 6 | 29.80 | scaffold1533 |  |  |  |  |  |  |  |  |  |  |
| ROS1 | M | R | 6 | 48.00 | ros\_assembly |  |  |  |  |  |  |  |  |  |  |
| ROS2 | M | R | 6 | 48.00 | ros\_assembly |  |  |  |  |  |  |  |  |  |  |
| ROS3 | M | R | 6 | 48.00 | ros\_assembly |  |  |  |  |  |  |  |  |  |  |
| ELUTA | M | R | 6 | 48.00 | ros\_assembly |  |  |  |  |  |  |  |  |  |  |
| VENOSA | M | R | 8 | 2.30 | scaffold901 |  |  |  |  |  |  |  |  |  |  |
| AS1\* | Y | S | 2 | 4.07 | scaffold455 |  |  |  |  |  |  |  |  |  |  |
| AS1\*\* | Y | S | 2 | 22.83 | scaffold530 |  |  |  |  |  |  |  |  |  |  |
| Am4'CGT | Y | S | 2 | 43.18 | scaffold316 |  |  |  |  |  |  |  |  |  |  |
| sulf | Y | R | 4 | 9.36 | scaffold91 |  |  |  |  |  |  |  |  |  |  |
| F3H | An/Fl | S | 5 | 20.44 | scaffold1246 |  |  |  |  |  |  |  |  |  |  |
| CHI | An/Fl | S | 2 | 41.99 | scaffold5972 |  |  |  |  |  |  |  |  |  |  |
| F3'5'H | An/Fl | S | 2 | 57.64 | scaffold260 |  |  |  |  |  |  |  |  |  |  |
| ANS | An/Fl | S | 2 | 73.07 | scaffold2068 |  |  |  |  |  |  |  |  |  |  |
| F3'H | An/Fl | S | 5 | 43.00 | scaffold609 |  |  |  |  |  |  |  |  |  |  |
| FLS | Fl | S | 5 | 32.32 | scaffold261 |  |  |  |  |  |  |  |  |  |  |
| CHS | FP | S | 4 | 65.82 | scaffold211 |  |  |  |  |  |  |  |  |  |  |

**Supporting Information**



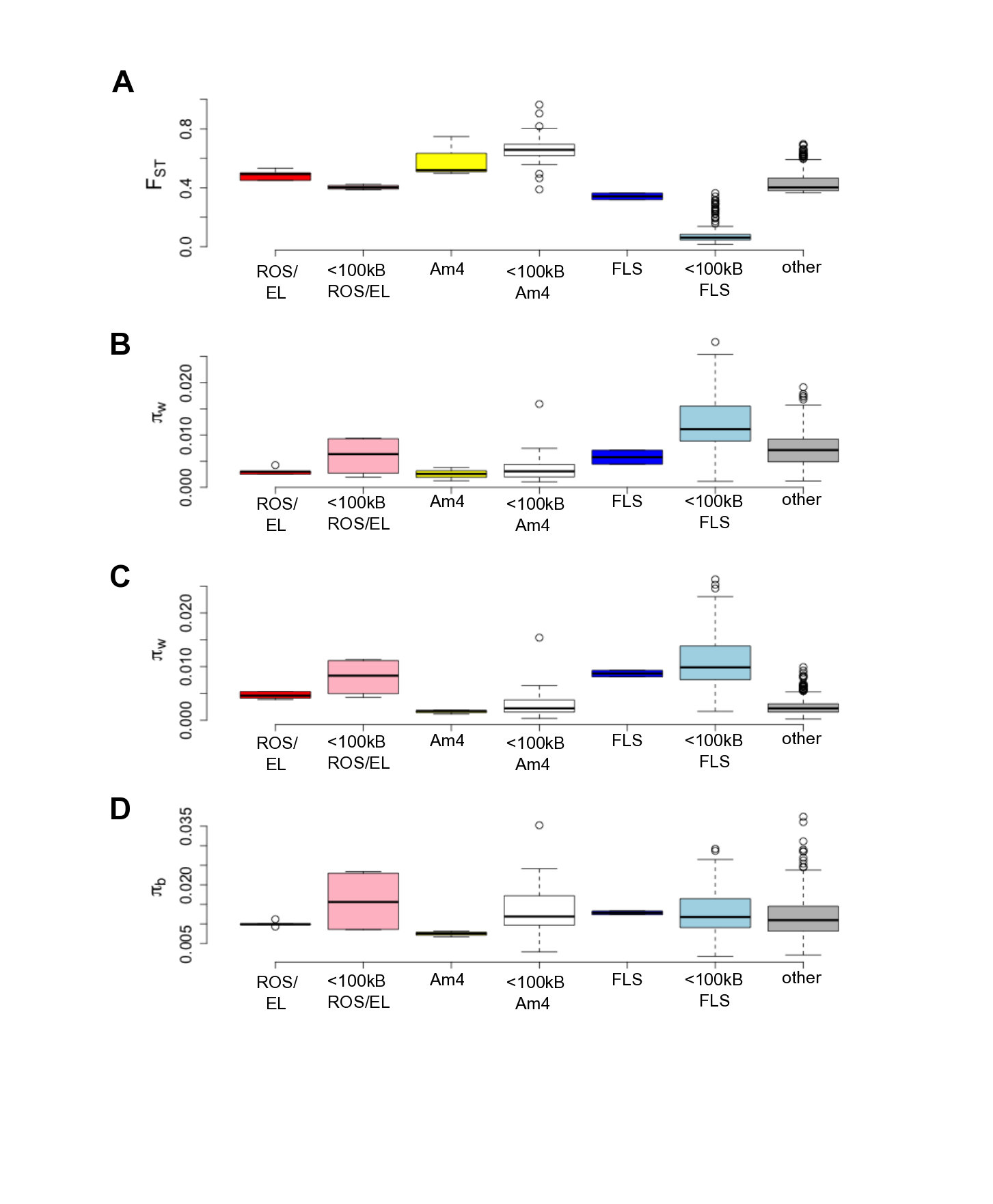
**Supporting Figure S1**. Relative divergence (FST) for 10kB windows along the eight linkage groups of *Antirrhinum majus*. The FST panels are for all *inter-subspecific* population pairs between *A. m. striatum* and *A. m. pseudomajus*. For each FST plot, windows are classified as background (grey), tightly linked (<300kB) to Am4’ CTG (yellow = downstream/orange = upstream), flavanol synthase FLS (blue) or the ROS/EL genes (red). Locations of flower colour genes also indicated as structural (triangle) or regulatory (diamond) with same colours as above if clines present and white if absent. Windows with excess relative divergence (>99.5th quantile) indicated with black lines above each panel (and dashed lines). Asterisk above panel A indicates genomic regions that were consistent outliers across multiple interspecific pairwise comparisons. Positions of windows and clines in physical distance (bp) along linkage groups.

****

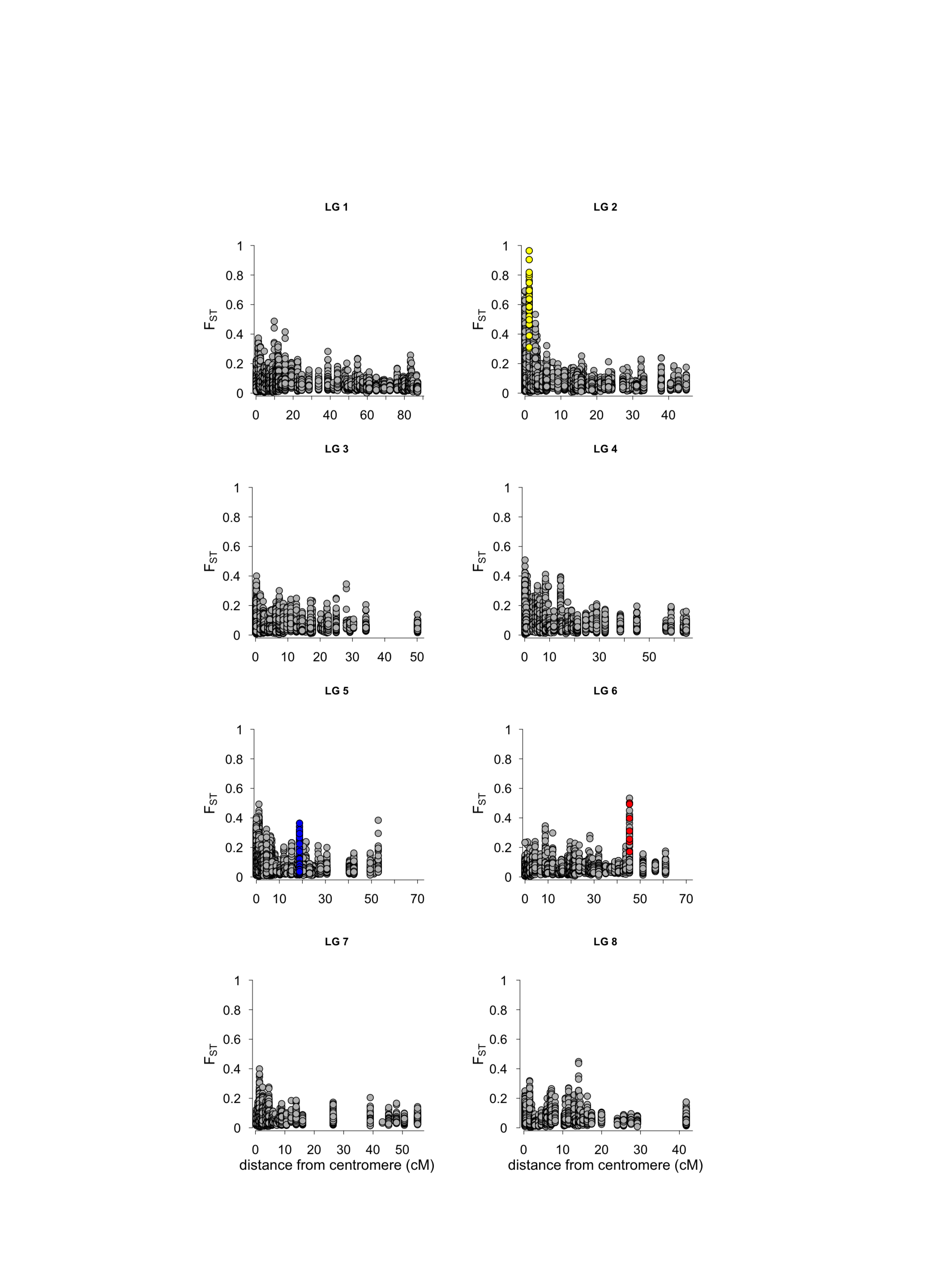
**Supporting Figure S2**. Relative divergence (FST) for 10kB windows along the eight linkage groups of *Antirrhinum majus*. The FST panels are for all *intra-subspecific* population pairs among *A. m. striatum* and *A. m. pseudomajus*. For each FST plot, windows are classified as background (grey), tightly linked (<300kB) to Am4’ CTG (yellow = downstream/orange = upstream), flavanol synthase FLS (blue) or the ROS/EL genes (red). Locations of flower colour genes also indicated as structural (triangle) or regulatory (diamond) with same colours as above if clines present and white if absent. Windows with excess relative divergence (>99.5th quantile) indicated with black lines above each panel (and dashed lines). Positions of windows and clines in physical distance (bp) along linkage groups.

Still to plot

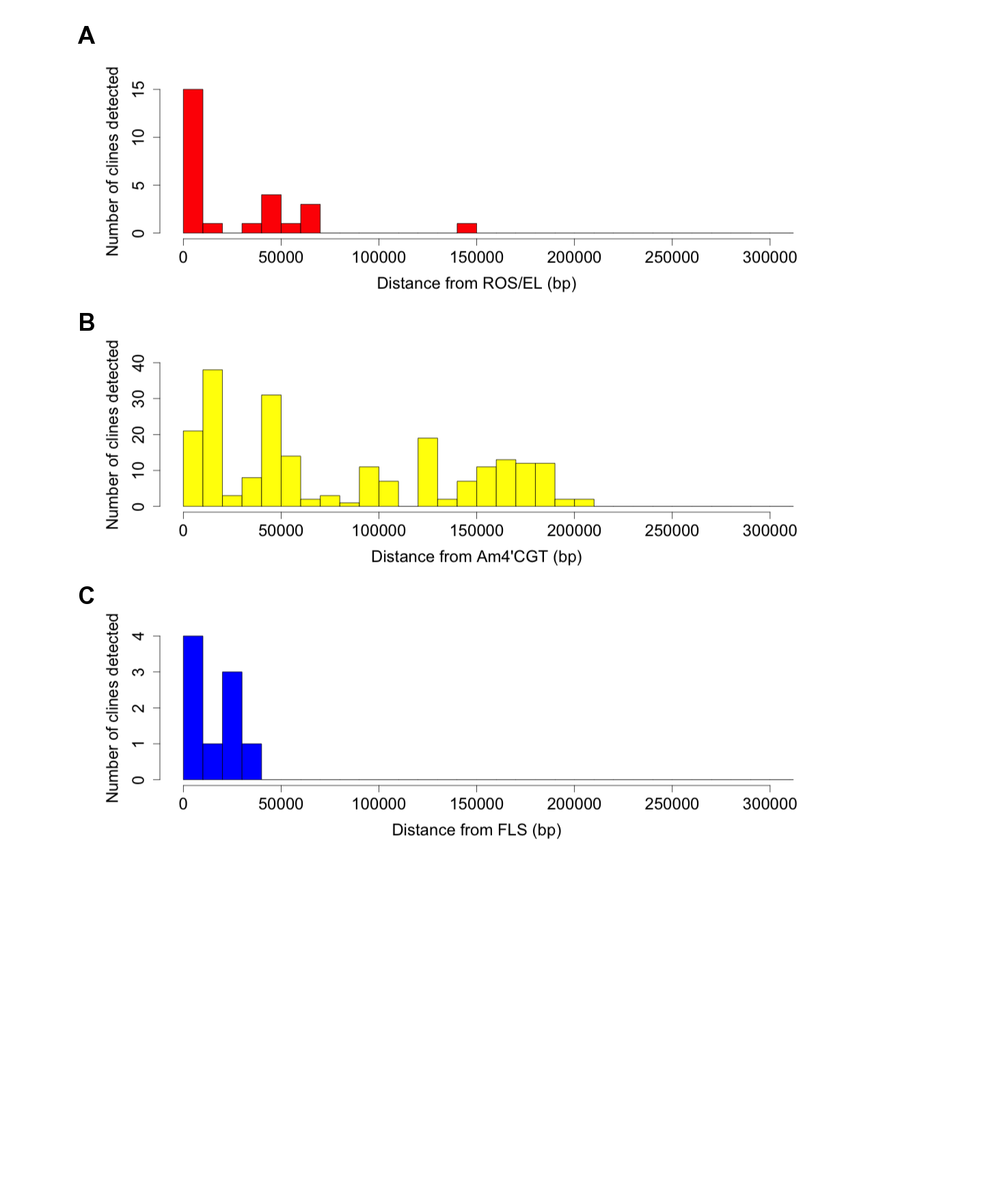
**Supporting Figure S3**. Pairwise divergence (*b*) for 10kB windows along the eight linkage groups of *Antirrhinum majus*. The panels are for all *inter-subspecific* population pairs among *A. m. striatum* and *A. m. pseudomajus*. For each plot, windows are classified as background (grey), tightly linked (<300kB) to Am4’ CTG (yellow = downstream/orange = upstream), flavanol synthase FLS (blue) or the ROS/EL genes (red). Locations of flower colour genes also indicated as structural (triangle) or regulatory (diamond) with same colours as above if clines present and white if absent. Windows with excess pairwise divergence (*b*) (>99.5th quantile) indicated with black lines above each panel (and dashed lines). Positions of windows and clines in physical distance (bp) along linkage groups.



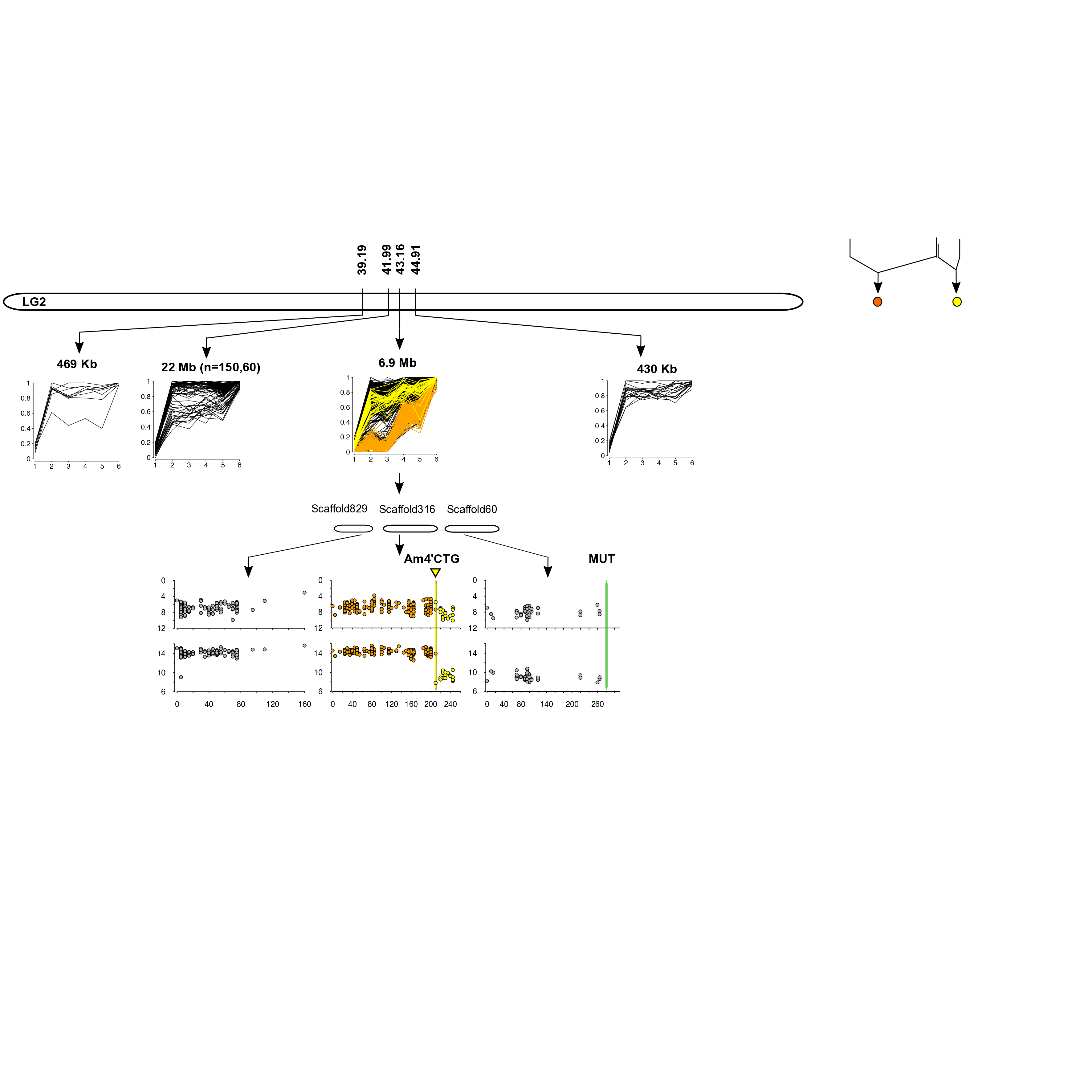
**Supporting Figure S4**. Comparison of relative divergence FST (A) nucleotide diversity within striatum (B) and pseudomajus (C) and absolute divergence (D) for the most distant population comparison (YP1 x MP6) for windows overlapping within known regulatory or structural flower colour genes (ROS/EL, Am4= Am4’CTG, FLS), and windows that are tightly linked to these genes (<100kB ROS/EL, Am4, FLS), upstream or downstream of Am4CTG (up str Am4, dw str Am4) or other regions > 500kB from known colour related genes.



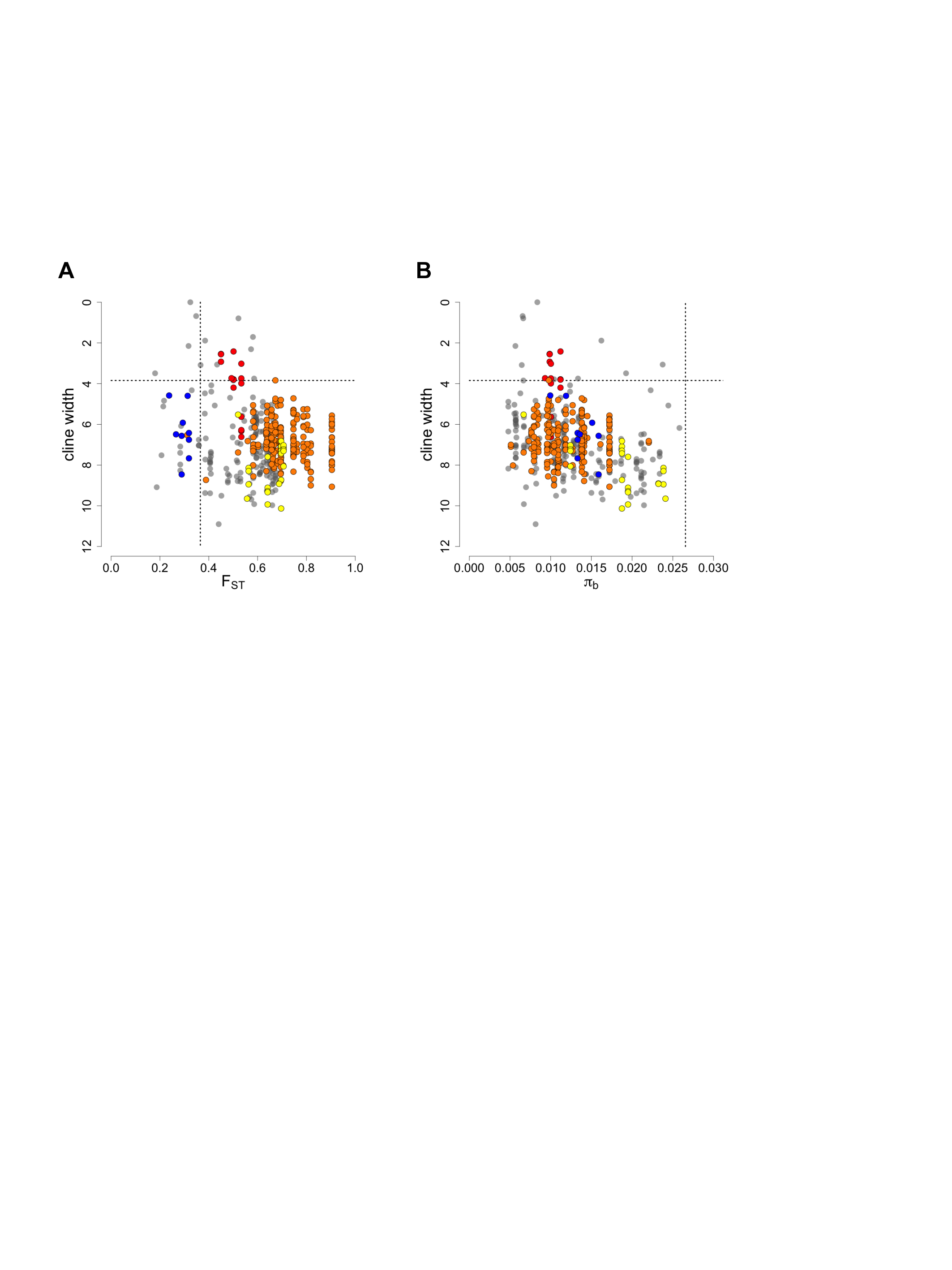
**Supporting Figure S5**. Approximate distance from centromere in relation to relative divergence (FST) for each of the eight linkage groups of *Antirrhinum*. For each plot, 10 kB windows are classified as background (grey), tightly linked (<300kB) to Am4’ CTG (yellow), flavanol synthase FLS (blue) or the ROS/EL genes (red).



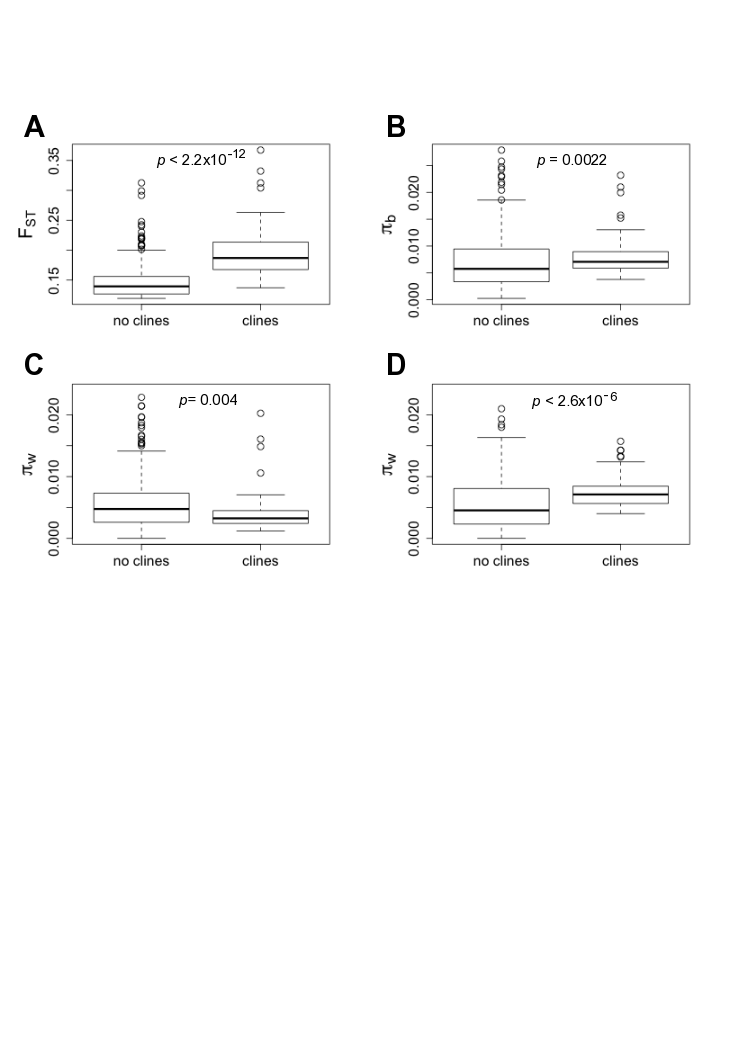
**Supporting Figure S6**. Number of clines detected in relation to distance from known flower colour genes (A) ROS/EL, (B) Am4’CGT, (C) FLS.



**Supporting Figure S7**. Staggered clines either side of the structural gene Am4’CGT



**Fig S8.** Relation between cline width and (A) FST and (B) absolute divergence, for the most distant populations (YP1 x MP6). Relative divergence (Fst) and absolute divergence calculated in 10kB windows, whereas cline width calculated for individual loci (n = 404).



**Supporting Fig S9.** Comparative tests of 10kB windows harboring steep clines vs. those without clines for the closer interspecific pairwise comparison on the same side of the mountain pass (YP2 x MP5). Measures of relative divergence FST (A), average polymorphism between populations, *b* (B), and polymorphism within *w* for striatum (C) and pseudomajus (D). Tests for significant differences from non-parametric Mann-Whitney U test.

**Supporting information S1**

*Geographic cline approximation: power analysis*

We used simulations to examine the power of the approximation method to distinguish coarse shifts in the centre and width of sigmoid clines. The position and width of clines is expected to vary from the true parameters due to drift that affects allele frequencies, binomial sampling of genotypes, sampling error, variation in sequencing depth and sequencing error. We modeled each of these factors using parameters similar to the whole genome data for *Antirrhinum* (i.e. 50 individuals pooled from each of eight demes). For all simulations we used eight demes in fixed locations, equally spaced 100 metres apart. The expected allele frequencies across the demes followed a logistic cline,

where *p0* and *p1* represent the allele frequencies in the parental demes either side of the cline (i.e. deme 1 and deme 8), *c* and *w* are the centre and width, and *x* a given position in space.

In each deme, the observed allele frequency fluctuates around the expected due to drift following a Beta distribution,

where Fst = 0.03, following estimates at neutral loci for *Antirrhinum* (this paper). Diploid genotypes were randomly drawn from the observed allele frequencies following a binomial and departures from random mating expectations at Hardy-Weinberg equilibrium (HW),

generating 50 individual genotypes in each deme. To simulate variation in sequencing depth, we used a normal distribution with a mean and variance following the observed values in the *Antirrhinum* aligned whole genome data [pool 1 and 6 , pool 2,3,4,5 ]

The cline centre and width was varied in steps of 100 metres [c={250,350,450,550,650}, w={50,150,250,350,450}] and for each set of parameter values we ran replicate simulations (*n*= 5000). To examine the affect of departures from HW, we ran simulations for one set of cline parameters (c = 450, w = 150) for a range of F = {0,0.5,0.1,0.2}. Similarly, to examine the effect of increased sequencing depth, we also ran this set of cline parameters for a higher depth .

**Supporting information S2. Cline BLAST**

While flower colour is implicated as the most significant reproductive barrier, clines elsewhere in the genome may signal uncharacterized loci (including other flower colour variants) that also contribute to reproductive isolation. Here, we examine the functional genes present at all divergent loci and asked whether parameters estimates for geographic and genomic clines are significantly different between loci linked to flower colour compared to other functional gene classes. The reference *A.* *majus* genome was used to examine sequence similarity between genomic regions possessing steep clines to annotated gene databases for *Antirrhinum* and angiosperms. This analysis was restricted to divergent regions with the strongest evidence of both high relative differentiation (*F*ST) and the presence of steep clines. Sites were included if they exhibited fixed or strong allele frequency differences between the outer populations ( > 0.9) that were significant after controlling for depth variation (Fischer Exact Test). The first 100 steepest clines were investigated by blasting xxx bp sequence 5’ and 3’ ends of the focal SNP to the draft *Antirrhinum* annotated genome (RNAseq transcriptome) and Angiosperm UniProt database. Genes and loci were broadly categorised into one of six groups based on gene product function: (i) color related genes, (ii) RNA/DNA, (iii) growth & metabolism, (iv) fatty acids, (v) retro elements and uncharacterized. Only genes with significant E-values and bit scores were included in this group system. In some cases, clines are most closely linked to one functional gene family whilst also being tightly linked to a flower colour gene. Given previous studies have shown steep clines may persist around flower colour genes up to 0.5cM, this makes it difficult to determine the driver. Considering the importance of flower colour in reproductive isolation, we classified loci in two ways, (i) based on the closest annotated gene, (ii) re-assigned as colour linked if < 0.5cM from known flower colour gene. We then simply compare the mean point estimates of cline parameters between the functional classes of genes exhibiting steep clines across the genome.

*BLASTs to the UniProt data base*

BLASTs to the UniProt Plants data base (2015-10 release) were conducted with an E-threshold of 10, Auto matrix settings (to accommodate the different BLAST sizes), no filtering, gapped allowance. When the BLASTs were conducted, 3,401,026 sequences containing 1,124,950,370 letters were contained within the data base.

Generally, BLASTs were conducted on the SNP associated with the high-cline gradient, as well as roughly 1000 surrounding base pairs (roughly 500 on either side). This size changed to accommodate annotation information. For instance, if the SNP was located in a gene that appeared in the annotation files, then the entire gene sequence as highlighted in the annotation was BLASTed to the database. Likewise, if the SNP was located in an LTR, the entire LTR (either larger or smaller than 1000bp) was BLASTed to the database. If the SNP was located next to an LTR, the BLASTed sequence purposely avoided containing sequence from the LTR.

BLAST size (column name: BLAST size) was recorded, as was whether an annotated gene was contained on or near the SNP (column name: Annotation 1). Only the closest gene to the SNP was included in the Annotation 1 column. (Next to the Annotation 1 column, a confidence score contained in the annotation file was also included [column name: Confidence]). Similarly, if the SNP was within or close to an LTR, the location of the LTR is recorded in the repeat masker annotation column. If the BLAST size seems unusually small or large, these columns may inform why. BLAST size may also have increased if multiple SNPs with high clines (uninterrupted by LTRs) were next to each other on a scaffold (usually, within 1000bp of each other).

If the BLAST returned a hit, information on the hit was recorded, including the name of the gene/protein, the Entry Name of the UniProt sequence that had sequence similarity, the E value, the score, and the Identity of the sequence. These three statistics can help determine whether the BLAST hit can be trusted or should be disregarded. If no hit was returned, “no hit” was recorded in the column.

If multiple entries were returned, the entry with the highest E-value was recorded. However, if this hit was “uncharacterized”, the hit with the highest (significant) E-value with a gene/protein name was recorded instead.

*BLAST Statistics*

E-Value: E-value informs statistical significance of BLAST hit. If E<10E-10, false positives for the overwhelming majority of sequences are eliminated. Using this conservative cut off, one can be relatively certain that the BLAST hit is correct, however a chance exists that some genes will be missed. Nevertheless, E<10E-10 should be used as the cutoff. (Pearson, 2013)

Bit Score: After eliminating hits based on E-value, bit score should be consulted. A bit score > 50 should be considered significant for databases with fewer than 7 million entries. Above 50, sequences are likely homologous. (Pearson, 2013)

Identity: Identity measures whether exactly the same nucleotide or amino acid is in the same position in the aligned sequence. This measure is not useful for inferring homology, however if the E-value and the bit score are significant, identities over 40% are likely to be orthologs based on enzyme commission numbers. (Pearson, 2013)

Pearson, William R. “An Introduction to Sequence Similarity (“Homology”) Searching”. Curt Proton Bioinformatics. 2013.

*Antirrhinum BLASTs*

BLASTs were also performed against a manually curated database consisting of ~80,000 antirrhinum scaffold sequences. *Antirrhinum* species sequences were then pulled from the UniProt database and BLASTed to this database. Confidence scores can be found in the LinkageMap document. Any scaffold that contained one of these genes received an annotation in the Antirrhinum column. For scaffolds with many high-cline SNPs, the *Antirrhinum* BLAST information was placed in columns for SNPs that were closest to the gene. This was to avoid unnecessarily cluttering the spread sheet, particularly for regions of Monster.

*Annotations*

As noted above, annotations from TriAnnot were incorporated into the spreadsheet and UniProt BLASTing criteria.

*Miscellaneous*

1. Genes with a known relationship to color determination were highlighted in yellow.
2. Genes were also grouped in to 5 categories based on gene product function: color, RNA/DNA, growth & metabolism, fatty acids, uncharacterized and retro elements. Only genes with significant E-values and bit scores were included in this group system.
3. glucosyl and glycosyl transferases have high sequence identity, so it is possible that a glycosyl transferase (important in the chalcone-yellow pathway) could be flagged as glucosyl.

**Supporting information S3.**

**The Importance of UDP-Glucose in determining *Antirrhinum* flower color**

UDP-Glucose is important in the synthesis of both aurone (yellow) and anthocyanin (red) pigmentation in flowers. Glycosyltransferases are key enzymes in the pathways that create these pigments, and these enzymes require a glycosyl donor. UDP-glucose is the native sugar donor for enzymes like 4’CGT, UFGT, and 5GT in antirrhinum.

Enzymes that require UDP-glucose include:

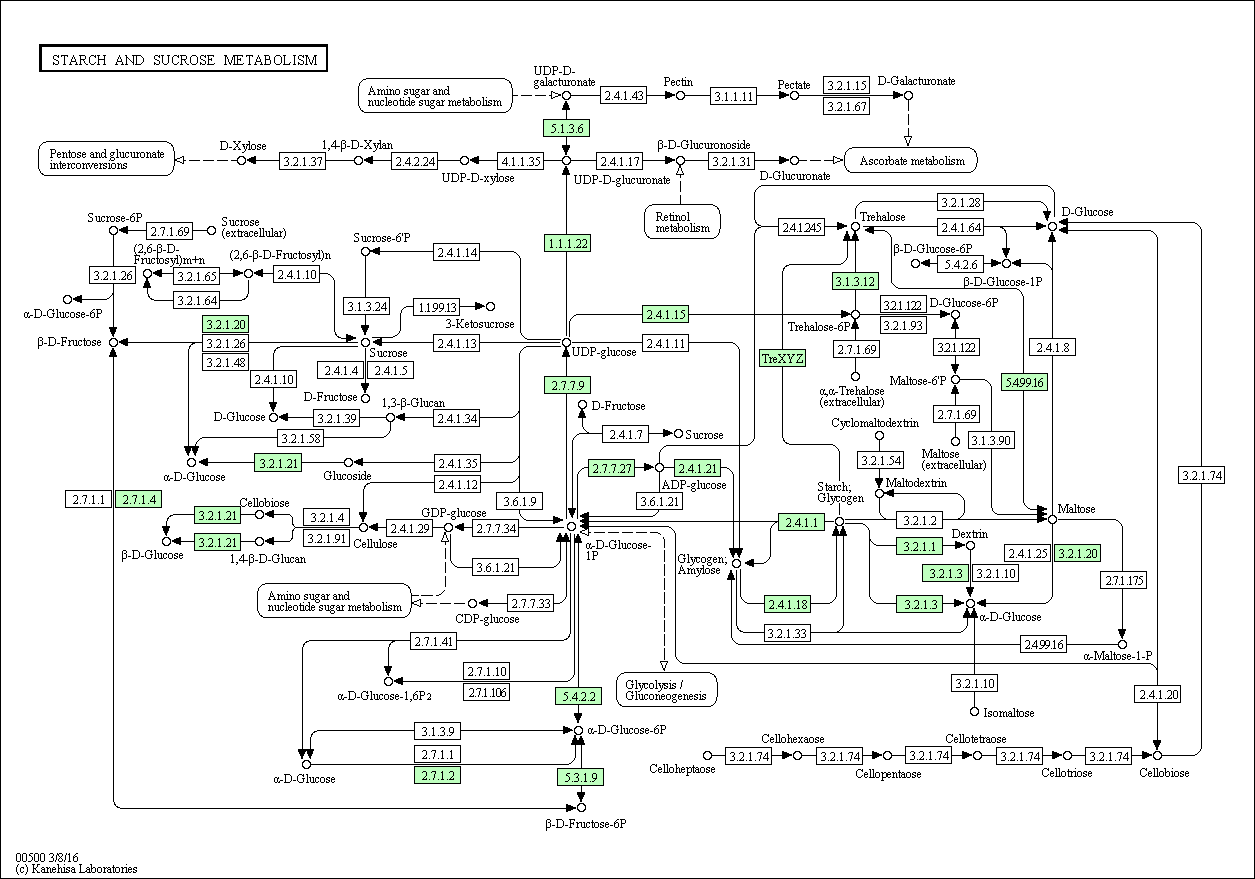
*Aurone (Yellow)*

* 4’CGT (UDP-glucose:chalcone 4’-O-glucosyltransferase)
  + Rxn: UDP-alpha-D-glucose + 2',3,4,4',6'-pentahydroxychalcone = UDP + 2',3,4,4',6'-pentahydroxychalcone 4'-O-beta-D-glucoside
  + UDP-alpha-D-glucose + 2',3,4,4',6'-pentahydroxychalcone = UDP + 2',3,4,4',6'-pentahydroxychalcone 4'-O-beta-D-glucoside

*Anthocyanin (magenta)*

* UFGT (UDP-glucose:flavonoid-3-O-glucosyltransferase)
* 5GT (UDP-glucose:anthocyanin 5’O’glucosyltransferase)
* UDP-glucose:anthocyanin 3’-O-glucosyltransferase
* UDP-glucose:anthocyanin 3’,5’-O’glucosyltransferase

Notably, 4’CGT has specificity for UDP-glucose (Ono et al., *PNAS*, 2006).

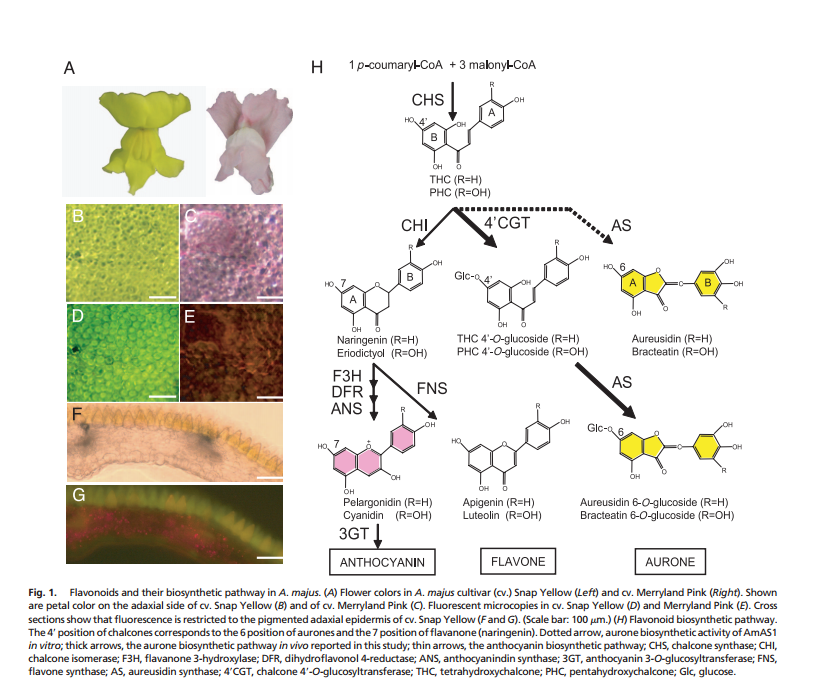


As demonstrated in the above Kegg Pathway, UDP-glucose is a hub in starch and sucrose metabolism. Some highlights of the function of UDP-glucose that this pathway illustrates include:

* Precursor for sucrose synthesis
* Precursor for cell wall polysaccharides (UDP-glucose by cellulose synthase and glucose-1-phosphate by ADP-glucose pyrophosphorylase)
* UDP-glucose converted to UDP-glucoronate by UDP-glucose dehydrogenase. UDP-glucoronate is a precursor for hemicellulose and pectin.
* Entry into the pentose phosphate pathway by glucose-6-phosphate (converted from glucose-1-phosphate by phosphoglucomutase)

Although UDP-glucose is important for a variety of pathways, as the glycosyl donor in pigment creation, it is a key metabolic limitation in color determination (Yan et al., *Biotechnology and Bioengineering*, 2007). Because UDP-glucose and its upstream substrates are situated at metabolic hubs, each molecule has the potential to enter into a number of pathways. Experiments in *E. coli* have shown that by manipulating specific enzymes, UDP-glucose can be synthesized intracellularly at higher rates. Mao and colleagues (*Biotechnol. Prog.*, 2006) demonstrated that overexpressing phosphoglucomutase (*pgm*)and UDP-glucose pyrophosphorylase (*gal*U) increases carbon flux toward UDP-glucose synthesis. This strategy targeted the branch point of glucose-6-phosphate. Building upon this technology, Yan and colleagues amplified *pgm* and *gal*U in concert with the flower pigment genes ANS and 3GT in genetically modified *E. coli*. By doing so, they increased pigment production by 57.8%, demonstrating that increased UDP-glucose formation is sufficient to increase pigment.

In *Antirrhinum*, both aurone (yellow) and anthocyanin (red) synthesis require UDP-glucose in catalysis of pigmentation reactions. Within this context, it is interesting to consider that phosphoglucomutase is located roughly 100kb upstream of a steep genomic cline. This cline occurs in a separate linkage group from other steep-cline color genes like ROS, EL, and Flavia (4’CGT). Scaffold1187 is in linkage group 1, at 2.17475 cM; it is not paracentromeric. Both ROS and EL feature high-cline SNPs within their genes, and while Flavia does not, Flavia occurs in an area of low recombination. 1187 may be paratelomeric. It is possible that an upstream regulatory mechanism affects *in vivo pgm* expression or phosphoglucomutase activity. However, because UDP-glucose is necessary in both aurone and anthocyanin pathways, it is unclear how increasing the concentration of UDP-glucose would subsequently drive one color production and not the other. It is possible that one pathway uses UDP-glucose sooner than the other, driving expression of one color more than the other.



From Ono et al., *PNAS*, 2006

As demonstrated above, chalcone synthase (CHS) begins both pathways. To reach anthocyanin pigment, five subsequent enzymes must catalyze reactions, with 3GT (and the donation of glucose by UDP-glucose) being the last. In contrast, in the production of Aurone pigment, 4’CGT catalyzes a reaction directly after CHS, moving the donation of glucose comparatively forward by four steps. Alterations in UDP-glucose availability could drive pigmentation toward or away from aurone or anthocyanin, although under this hypothesis, and increase in UDP-glucose would drive an increase in aurone (yellow) pigmentation. Alternatively, shunting UDP-glucose away from pigmentation pathways could also create more aurone pigmentation through the AS pathway that does not use glucose. (To test this hypothesis, it would be interesting to see if the different pathways create different shades of yellow and to see which is more abundant). This would require a delicate balance of enough UDP-glucose for cells to function, but not enough for use in pigmentation. Such a balance seems like it would be more difficult to create than an excess, but evolution has come up with crazier things.