Appendices to manuscript entitled:

Using neutral cline decay to estimate contemporary dispersal: a generic tool and its application to a major crop pathogen

A. RIEUX *, T. LENORMAND ¶, J. CARLIER *, L. DE LAPEYRE DE BELLAIRE † and V. RAVIGNE *

Appendix S1: Correspondence between sampling sites in Rieux et al. (2011) and those considered in the present study.

Appendix S2: Inference of the number of cluster(s) (k) and assignment probabilities using GENELAND and TESS.

Appendix S3: Summary of basic genetic diversity results.

Appendix S4: Finding the best cline shape.

Appendix S5: Likelihood ratio tests for locus/locus vanishing cline patterns test.

Appendix S6: Details on the model used for estimation of σ.

Appendix S7: Likelihood profile for the estimation of σ.

Appendix S8: Examples of kurtosis values found in the literature.

Appendix S9: Dispersal kernel families investigated using simulations.

References cited
Appendix S1: Correspondence between sampling sites in Rieux et al. (2011) and those considered in the present study.

Figure S1. Graphical representation of the sampling sites from Rieux et al. (2011) that have been also considered in the present study ($T_1$ sampling).
Appendix S2: Inference of the number of cluster(s) (K) and assignment probabilities using GENELAND and TESS.

Table S2. Inference of the number of cluster(s) (K) that best explains observed spatial genetic structure. For GENELAND we reported the percentage of iterations spend by the RJMCMC on each K across simulations after the burning period. For TESS we reported the values of DIC (Deviation Index Criterion), smallest values reflecting the best number of parental populations. It is noteworthy that it is not possible to run TESS for K = 1 (as a test of no population structure). Both indices were averaged over 5 runs with standard deviation indicated inside brackets. T₁ holds for first sampling, and T₂ for second sampling.

<table>
<thead>
<tr>
<th>K</th>
<th>GENELAND T₁</th>
<th>GENELAND T₂</th>
<th>TESS T₁</th>
<th>TESS T₂</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.31 (4.81)</td>
<td>2.16 (1.03)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>69.26 (3.46)</td>
<td>61.21 (5.72)</td>
<td>2097.13 (7.45)</td>
<td>3687.68 (6.22)</td>
</tr>
<tr>
<td>3</td>
<td>15.41 (1.29)</td>
<td>28.33 (2.75)</td>
<td>2152.54 (12.07)</td>
<td>3703.48 (9.01)</td>
</tr>
<tr>
<td>4</td>
<td>3.12 (1.87)</td>
<td>4.56 (3.77)</td>
<td>2251.29 (4.77)</td>
<td>3733.62 (8.77)</td>
</tr>
<tr>
<td>5</td>
<td>0.16 (0.93)</td>
<td>1.35 (0.44)</td>
<td>2410.38 (8.74)</td>
<td>3819.09 (3.75)</td>
</tr>
<tr>
<td>6</td>
<td>0.97 (1.63)</td>
<td>2.09 (0.99)</td>
<td>2382.94 (1.01)</td>
<td>4053.82 (12.99)</td>
</tr>
<tr>
<td>7</td>
<td>1.23 (1.04)</td>
<td>0.03 (0.17)</td>
<td>2389.06 (3.64)</td>
<td>4156.39 (1.48)</td>
</tr>
<tr>
<td>8</td>
<td>2.73 (0.98)</td>
<td>0 (0)</td>
<td>2508.51 (5.75)</td>
<td>4038.44 (3.69)</td>
</tr>
<tr>
<td>9</td>
<td>0 (0)</td>
<td>0.27 (1.39)</td>
<td>2592.95 (1.47)</td>
<td>4216.53 (5.78)</td>
</tr>
<tr>
<td>10</td>
<td>0.81 (0.57)</td>
<td>0 (0)</td>
<td>2442.72 (3.11)</td>
<td>4202.76 (6.74)</td>
</tr>
</tbody>
</table>
**Figure S2.** Assignment probability of either each individual to one of the two clusters (GENELAND) or each individual genome to one of the two parental populations (TESS) at both sampling times ($T_1$ and $T_2$). Each point represents an individual. The difference in number of points between GENELAND and TESS is due to the fact that GENELAND displays an average probability of assignment over individuals sampled at the same locality.
**Appendix S3: Summary of basic genetic diversity results**

<table>
<thead>
<tr>
<th>Locus</th>
<th>$T_1$ (N=144)</th>
<th>$T_2$ (N=244)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$Ar$</td>
<td>$H_E$</td>
</tr>
<tr>
<td>MISSR401</td>
<td>4</td>
<td>2.179</td>
</tr>
<tr>
<td>MISSR405</td>
<td>4</td>
<td>0.464</td>
</tr>
<tr>
<td>MISSR407</td>
<td>4</td>
<td>0.499</td>
</tr>
<tr>
<td>MISSR413</td>
<td>4</td>
<td>0.495</td>
</tr>
<tr>
<td>MISSR417</td>
<td>4</td>
<td>0.484</td>
</tr>
<tr>
<td>MISSR425</td>
<td>4</td>
<td>0.254</td>
</tr>
<tr>
<td>MISSR428</td>
<td>4</td>
<td>0.482</td>
</tr>
<tr>
<td>MISSR430</td>
<td>4</td>
<td>0.055</td>
</tr>
<tr>
<td>MISSR434</td>
<td>4</td>
<td>0.478</td>
</tr>
<tr>
<td>MISSR340†</td>
<td>2</td>
<td>0.352</td>
</tr>
<tr>
<td>MISSR362‡</td>
<td>2</td>
<td>0.12</td>
</tr>
<tr>
<td>MISSR322‡</td>
<td>2</td>
<td>0.563</td>
</tr>
<tr>
<td>MISSR324‡</td>
<td>2</td>
<td>0.455</td>
</tr>
<tr>
<td>MISSR350‡</td>
<td>2</td>
<td>0.188</td>
</tr>
<tr>
<td>MISSR203‡</td>
<td>3</td>
<td>0.402</td>
</tr>
</tbody>
</table>

**Table S3.** Summary of basic information and population genetic analyses of both temporal samplings ($T_1$ and $T_2$) computed for all microsatellite loci. $N$ is sample size. For each locus, we provide $H_E$, the unbiased estimate of gene diversity, $Ar$, the number of alleles and the $F_{ST}$ between the two GENELAND clusters with their significativity according to Fisher’s exact test (NS > 0.05; * P < 0.05; ** P < 0.01; *** P < 0.001). The number into brackets indicates the number of repeat unit nucleotides for each microsatellite locus. † Indicates the loci that were also at use in the previous study (Rieux et al. 2011). Nomenclature has changed between the two studies. MfSSR340, 362, 322, 324, 350 and 203 now respectively state for F40, F62, Fe05b, Fe09, FF09 and N203.
Appendix S4: Finding the best cline shape

Three different shapes were fitted to allelic count data.

- **Shape A** is called stepped cline and corresponds to a three-part function containing a central sigmoid and two exponential tails. Two parameters characterize the central sigmoid (center $c$ and slope $b$) and four parameters characterize the two tails (distance from the centre at which the right and left exponential tails start and their relative slopes, $d_1$, $d_2$, $t_1$ and $t_2$ respectively). Such a shape is expected in tension zones involving strong indirect selection across the genome (Szymura and Barton 1986).

  
  
  $1 - \frac{1}{1 + e^{-bd_2} e^{\frac{-bh_2(x-c-d_2)}{1+e^{bd_2}}}}$ for $x \leq c - d_2$

  
  
  $\frac{e^{b(x-c)}}{1 + e^{b(x-c)}}$ for $c - d_2 \leq x \leq c + d_1$  

  
  
  $1 - \frac{1}{1 + e^{-bd_1} e^{\frac{-bh_1(x-c+d_1)}{1+e^{bd_1}}}}$ for $x \geq c - d_1$

- **Shape B** is a simple Logit (two parameters). In such a shape, up and down asymptotic frequencies (plateaux) are automatically fixed to a value of 1 and 0 respectively (Szymura and Barton 1986). It is expected when when populations with diagnostic alleles meet (Endler 1977).

  
  
  $\frac{e^{b(x-c)}}{1 + e^{b(x-c)}}$  

  

The logit shapes naturally result from the resolution of the diffusion equation with the assumption of a linear secondary contact at time 0. These initial conditions assume that the two populations meet along a straight line and that:

  
  
  $\frac{\partial p(x,y,t)}{\partial t} = \frac{\sigma^2}{2} \nabla^2 p(x,y,t)$  

  

$p(x,y,0) = h_1$ for all $x,y$ on a side of the contact line and

  
  
  $p(x,y,0) = h_2 = h_1 + h \cdot (1 - h_1)$ for all $x,y$ on the other side of the contact line.

Note that the exact solution is in the form of an error function and that it is finely approximated using Logit fits.
• **Shape C** is a scaled Logit in which both up ($h_2$) and down ($h_1$) asymptotic frequencies might take any value between 0 and 1 (total of four parameters). It is expected in a contact zone with semi-diagnostic alleles (Endler 1977).

\[
h_1 + (h_2 - h_1) \frac{e^{b(x-c)}}{1 + e^{b(x-c)}}
\]

S4.4

To determine the best description of the data, we started by fitting those three different shapes to multilocus genetic data (15 Loci) independently for each temporal sampling ($T_1$ and $T_2$) and without setting any constraint onto parameters. Note that because we considered 2D versions of these functions,

\[
x = X \cos(\theta) + Y \sin(\theta)
\]

S4.5

where $X$ and $Y$ are sample longitude and latitude and $\theta$ the angle that the cline axis forms with the parallel. As a consequence, one supplementary parameter should be added to each of these three shapes to fit $\theta$, resulting in the existence of $k = 7, 3$ and 5 parameters for shape A, B and C respectively for each sampling date and locus (thus explaining the total numbers of parameters in table S4).

<table>
<thead>
<tr>
<th></th>
<th>Shape</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A</td>
<td>B</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>LL $T_1$</td>
<td>-1009.06</td>
<td>-1035.35</td>
<td>-993.10</td>
<td></td>
</tr>
<tr>
<td>LL $T_2$</td>
<td>-1651.15</td>
<td>-1716.77</td>
<td>-1629.26</td>
<td></td>
</tr>
<tr>
<td>LL Total</td>
<td>-2660.22</td>
<td>-2752.12</td>
<td>-2622.37</td>
<td></td>
</tr>
<tr>
<td>$Total k$</td>
<td>210</td>
<td>90</td>
<td>150</td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>5740.43</td>
<td>5684.24</td>
<td>5544.73</td>
<td></td>
</tr>
</tbody>
</table>

Table S4. Log likelihood (LL), total number of parameters ($Total k$) and corresponding AIC for the multilocus clines (15L) independently fitted for each sampling. Best value is highlighted in bold.
### Table S5. Tests of a vanishing cline pattern for each of the 15 microsatellite markers (MfSSR). Each locus was fitted using the scaled Logit shape. For each locus, a vanishing cline pattern was tested using likelihood ratio tests (LRT). LRTs were realized between the sum of likelihood of the two independent unconstrained models ($LL_{T_1+T_2} = LL_{T_1} + LL_{T_2}$) with the likelihood of a model in which both the slope and the asymptotic frequencies were constrained to be the same between $T_1$ and $T_2$ ($LL_0$). Significant values (5% level) are highlighted in bold (Dev holds for deviance). Numbers of parameters $k$ are in brackets.
Appendix S6: Details on the model used for the estimation of $\sigma$

In the present paper, clines were appropriately modelled using a scaled logit shape. At date $i$, the allelic frequency along the cline is:

$$p_i(x) = h_i + h_p(1-h_1) \frac{e^{b_i(x-x)}}{1+e^{b_i(x-x)}}$$  \hspace{1cm} S6.1

where $x$ is the position along the cline, $c_i$ is cline centre at date $i$, $b_i$ is a slope parameter (see below), $h_1$ is the lower asymptotic frequency and $h_p$ is the step between the lower and upper asymptotic allelic frequencies. Cline higher asymptotic frequency $h_2$ is parameterized as:

$$h_2 = h_i + h_p(1-h_1)$$  \hspace{1cm} S6.2

**Cline width of scaled logit clines**

Cline width is defined as the ratio of the maximal difference in allele frequencies along the cline and the maximal slope of the cline (i.e., slope at cline centre). So at date $i$, cline width is:

$$w_i \equiv \left| p_i(\infty) - p_i(-\infty) \right| / p'_i(c)$$  \hspace{1cm} S6.3

With the scaled logit shape,

$$\left| p_i(\infty) - p_i(-\infty) \right| = h_p(1-h_1)$$  \hspace{1cm} S6.4

and

$$p'_i(c) = \frac{b_i}{4} h_p(1-h_1)$$  \hspace{1cm} S6.5

Consequently, cline width is simply:

$$w_i = \frac{4}{b_i}$$  \hspace{1cm} S6.6

**Relationship between the change in cline shape and $\sigma$**

From Endler’s (1977) relationship:

$$w_i = \sigma \sqrt{2\pi T_i}$$  \hspace{1cm} S6.7
This translates into:

\[ T_i = \frac{w_i^2}{2\pi \sigma^2} \]  

S6.8

Therefore the number of generations between the two sampling events is:

\[ T = T_2 - T_1 = \frac{w_2^2 - w_1^2}{2\pi \sigma^2} \]  

S6.9

Using Equation S6.6, this becomes:

\[ T = \frac{8}{\pi \sigma^2} \left( \frac{1}{b_2^2} - \frac{1}{b_1^2} \right) \]  

S6.10

Solving for \( b_2 \), we get:

\[ b_2 = \frac{1}{\sqrt{\frac{1}{b_1^2} + \frac{\pi T \sigma^2}{8}}} \]  

S6.11

It is further considered that cline center moves by a step \( \delta \), such that:

\[ c_2 = c_1 + \delta \]  

S6.12

At date \( T_2 \), the allelic frequency along the cline is thus:

\[ p(x) = h_i + h_p (1-h_i) \frac{e^{b_2 (c_i + \delta - x)}}{1 + e^{b_2 (c_i + \delta - x)}} \]  

S6.13

**What the modified version of CFIT does**

For each locus in the dataset, CFIT fits equation S6.1 to the first sampling event and equations S6.11 and S6.13 to the second one using maximum likelihood. Such a parameterization allows obtaining locus-specific estimates of \( \sigma \) and testing hypotheses by constraining some parameters (\( c_1, b_1, \delta, \sigma \)) between loci or sampling events to be the same. In its current version, Cfit only allows estimating dispersal considering the scaled logit shape but it could be easily adapted to any other function.
Appendix S7: Likelihood profile for the estimation of $\sigma$

Figure S7. Likelihood profile for $\sigma$ (in m/generation$^{1/2}$). Dashed arrow indicates the 95% support limit interval around the likelihood maximum (black point) obtained from model G (see Table 2 for details on the different models).
### Appendix S8: Examples of kurtosis values found in the literature

<table>
<thead>
<tr>
<th>Scientific name</th>
<th>Commun name</th>
<th>Class</th>
<th>Kurtosis $\gamma$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Podisma pedestris</em></td>
<td>Grasshopper</td>
<td>Insect</td>
<td>$\sim 0$</td>
<td>Barton &amp; Hewitt 1982 (*)</td>
</tr>
<tr>
<td><em>Quercus lobata</em></td>
<td>Valley oak</td>
<td>Angiosperm</td>
<td>10.06 (*)</td>
<td>Auzterlitz et al, 2004</td>
</tr>
<tr>
<td><em>Dipodomys spectabilis</em></td>
<td>Kangaroo rats</td>
<td>Mammalia</td>
<td>19.4</td>
<td>Waser &amp; Elliott 1991 (*)</td>
</tr>
<tr>
<td><em>Sorbus torminalis</em></td>
<td>Wild service tree</td>
<td>Angiosperm</td>
<td>30.68 (*)</td>
<td>Auzterlitz et al, 2004</td>
</tr>
<tr>
<td><em>Puccinia striiformis</em></td>
<td>Wheat stripe rust causal agent</td>
<td>Urediniomycetes</td>
<td>47.2 (*)</td>
<td>Sackett &amp; Mundt 2005</td>
</tr>
</tbody>
</table>

Table S8. Examples of normalized kurtosis values found in the literature. The (*) indicates that we did not explicitly find the kurtosis value in the paper of reference but had to compute it from parameter values indicated in the study as $\gamma = \frac{\mu_4}{\sigma^4} - 3$.

In this equation, $\mu_4$ and $\sigma^2$ are respectively the fourth and the second moment around the mean.
Appendix S9: Dispersal kernel families investigated using simulations.

In addition to the use of the mixture of Binomial (equation can be found in Lenormand 1998), we assessed the robustness of the estimation of $\sigma$ using three other families of functions commonly used to model dispersal kernels. These are the, Exponential power, Clark’s 2Dt and Geometrical distribution families (equations can be found in Klein et al. 2006). Equations from Klein et al. 2006 were modified such as the scale parameter sums at 1 (because of the 1D nature of our simulations). $\Gamma$ stands for the gamma function such as:

$$\Gamma(x) = \int_0^{+\infty} t^{x-1} e^{-t} dt$$  \hspace{1cm} S9.1

**Mixture of Binomial distribution family**

$$m(x) = \alpha_1 B(2\mu_1, \frac{1}{2}) + \alpha_2 B(2\mu_2, \frac{1}{2})$$  \hspace{1cm} S9.2

with $B$ holding for Binomial distribution

**Exponential power family:**

$$m(x) = \frac{\beta}{2\varepsilon \Gamma(1/\beta)} \exp\left(-\left(\frac{|x|}{\varepsilon}\right)^\beta\right)$$  \hspace{1cm} S9.3

with $\beta > 0$, $\varepsilon > 0$

**Geometric family:**

$$m(x) = \frac{\beta - 1}{2\varepsilon} \left(1 + \frac{|x|}{\varepsilon}\right)^{-\beta}$$  \hspace{1cm} S9.4

with $\beta > 5$ and $\varepsilon > 0$.

**Clark’s 2Dt family:**

$$m(x) = \frac{\Gamma(\beta)}{\sqrt{\pi} \Gamma\left(\beta - \frac{1}{2}\right)} \left(1 + \frac{x^2}{\varepsilon}\right)^{-\beta}$$  \hspace{1cm} S9.5

with $\beta > 5/2$ and $\varepsilon > 0$. 


