On the usefulness of genealogical trees

N. Barton, A.M. Etheridge, A. Véber and friends.

CIRM - 13/06/2012

Evolution in a continuum

Aim: Model the evolution of the genetic composition of a geographically structured population. Space is continuous (and in 2 dimensions, most of the time).

Main characteristics

- Reproduction happens more or less locally;
- At ‘stationarity’, local population sizes are regulated;
- Individuals have a finite pool of potential parents ($\Rightarrow$ multiple mergers in the genealogies);
- Rare but severe bottlenecks can occur and affect potentially large regions.

Main characteristics

- Reproduction happens more or less locally;
- At ‘stationarity’, local population sizes are regulated;
- Individuals have a finite pool of potential parents (⇒ multiple mergers in the genealogies);
- Rare but severe bottlenecks can occur and affect potentially large regions.

Questions of interest

- Behaviour under the hypothesis of neutrality?
- Spatial decay of correlations between local genetic diversities?
- Signature of a deviation from “*local rep. + neutrality*”?
  - ↪ large but rare extinction/recolonisation events;
  - ↪ selection and selective sweeps;

And also...

- Which are the quantities summarizing the evolution?
- How can we infer them from data?
- Can we detect deviations from neutrality?

Remarks

- Already well-studied: Wright's *island model*, the *stepping-stone model*.
  - We shall obtain equivalent results in continuous space, under equivalent assumptions;
  - But we can accommodate many other scenarii than the ‘classical ones’.

- For the rest of the talk, imagine a population of plants.

An event-based model

- Fix $\lambda > 0$ and a measure $\xi(dr, du)$ on $(0, \infty) \times [0, 1]$.
- **Reproduction events:** given by a Poisson point process on $[0, \infty) \times \mathbb{R}^2 \times (0, \infty) \times [0, 1]$ with intensity measure $dt \otimes dx \otimes \xi(dr, du)$.

In words, we define a random sequence $\{(t_i, x_i, r_i, u_i),\ i \in \mathcal{I}\}$ of times, centres, radii and impacts.

An event-based model

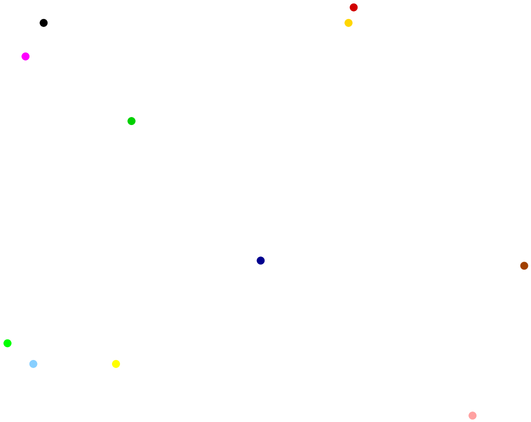
- Fix $\lambda > 0$ and a measure $\xi(dr, du)$ on $(0, \infty) \times [0, 1]$.
- **Reproduction events:** given by a Poisson point process on $[0, \infty) \times \mathbb{R}^2 \times (0, \infty) \times [0, 1]$ with intensity measure $dt \otimes dx \otimes \xi(dr, du)$.

In words, we define a random sequence $\{(t_i, x_i, r_i, u_i),\ i \in \mathcal{I}\}$ of times, centres, radii and impacts.

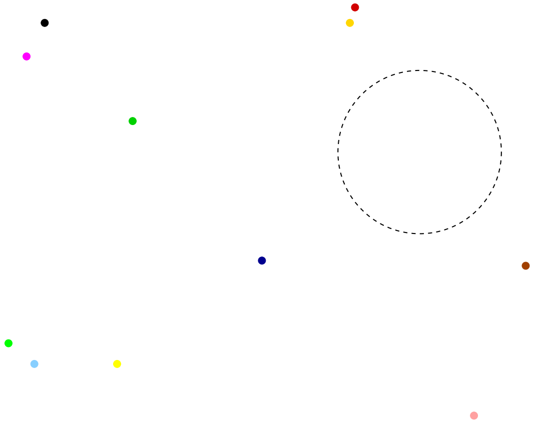
We start from a Poissonian cloud of indv. At the time t_i of an event, if $B(x_i, r_i)$ is empty, then do nothing. Otherwise, within the ball

1. Choose a parent uniformly at random;
2. Each indv. within the ball dies with proba u_i, indep. of each other;
3. Add a Poissonian cloud of new indv. with density λu_i. All of them have the same allele as the parent.

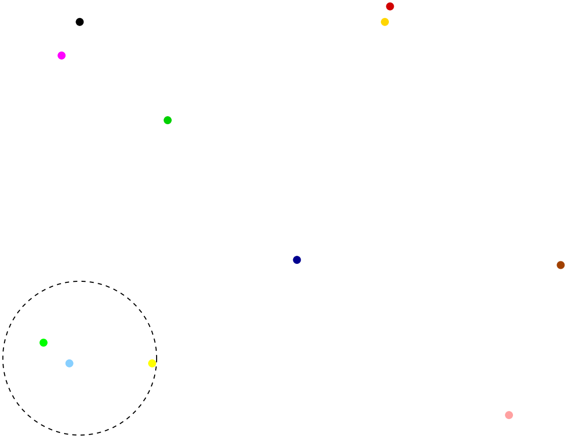
In pictures



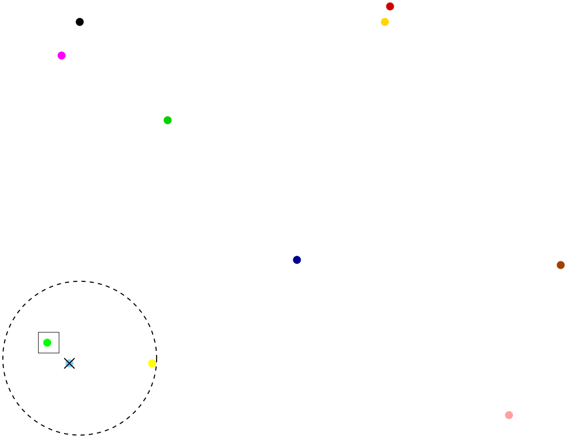
In pictures



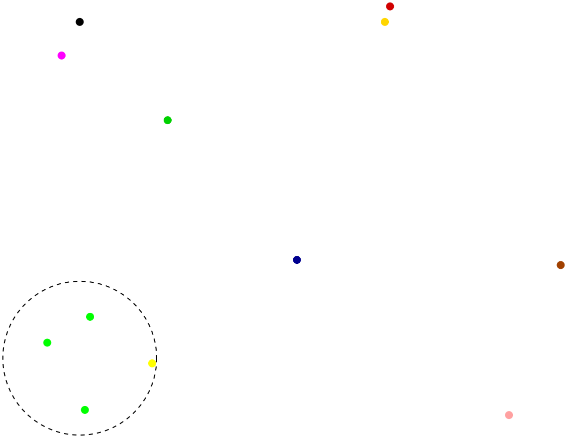
In pictures



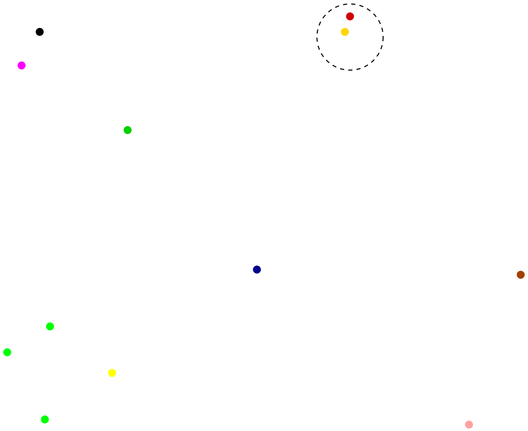
In pictures



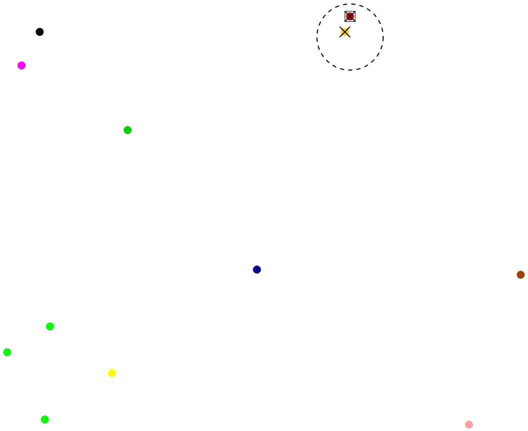
In pictures



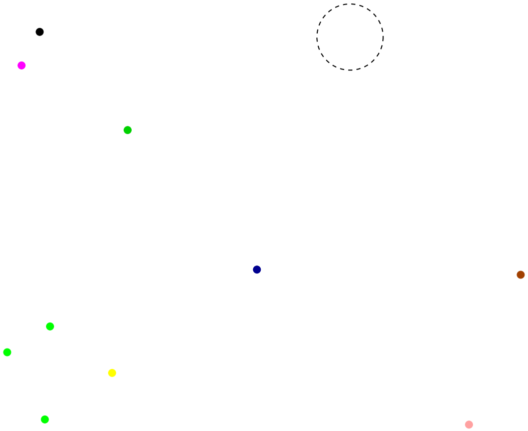
In pictures



In pictures



In pictures



A few comments

- **Objectives met:** In a populated region, each individual reproduces rarely $\Rightarrow$ sort of *logistic* regulation. Other characteristics included as well.
- **A flexible framework:** replace the balls by Gaussian kernels, or any mechanism preserving the average local density of indv.
- **Berestycki, Etheridge & Hutzenthaler (2010):** If λ is large enough, the population survives and has a stationary distribution.
- **But:** Genealogies are not easy to describe, since the presence of an individual gives us information on the past (not a simple time reversal). Forwards-in-time model not very tractable either.

To cope with the last issue, we let the density λ tend to infinity.

$\Rightarrow$ In the limit, the population covers the whole plane $\mathbb{R}^2$.

The spatial Λ-Fleming-Viot process

Type/allele space : K compact.

Population at time t **:** Measure M_t on $\mathbb{R}^d \times K$ whose first marginal is Lebesgue measure (uniform density of indv.). That is,

$$M_t(dx, dk) = dx\, \rho_t(x, dk).$$

A possible interpretation: The ‘real’ population is a Poisson point process with (random) intensity measure M_t (Wakolbinger & V., 2012).

The spatial Λ-Fleming-Viot process

Type/allele space : K compact.

Population at time t : Measure M_t on $\mathbb{R}^d \times K$ whose first marginal is Lebesgue measure (uniform density of indv.). That is,

$$M_t(dx, dk) = dx\, \rho_t(x, dk).$$

A possible interpretation: The 'real' population is a Poisson point process with (random) intensity measure M_t (Wakolbinger & V., 2012).

Evolution : same Poisson point process of events. If t_i is the time of an event, the reproduction event occurs within $B(x_i, r_i)$.

- A parent is chosen uniformly at random from $B(x_i, r_i)$ [*location z, type* κ];
- For every $y \in B(x_i, r_i)$, $\quad \rho_{t_i}(y, dk) = (1 - u_i)\rho_{t_i-}(y, dk) + u_i\delta_\kappa$.

Duality relations

- The genealogical process $(\{\xi_s^1, \ldots, \xi_s^{N_s}\})_{s\geq 0}$ is a system of *a priori* correlated (symmetric) jump processes that coalesce when they are *affected* by the same event.

- Take $K = \{0, 1\}$ and $w_t(x) := \rho_t(x, \{1\})$. Then, we have: for every $j \geq 1$ and $\psi \in C_c((\mathbb{R}^d)^j)$,

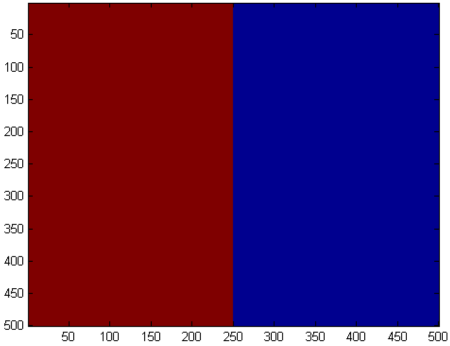
$$
\begin{aligned}
&\mathbb{E}_{w_0}\left[\int_{(\mathbb{R}^d)^j} \psi(x_1, \ldots, x_j)\left\{\prod_{i=1}^{j} w_t(x_i)\right\} dx_1 \cdots dx_j\right] \\
&\qquad = \int_{(\mathbb{R}^d)^j} \psi(x_1, \ldots, x_j)\mathbf{E}_{\{x_1,\ldots,x_j\}}\left[\prod_{i=1}^{N_t} w_0\left(\xi_t^i\right)\right] dx_1 \cdots dx_j.
\end{aligned}
$$

In particular,

$$
\mathbb{E}_{w_0}\left[\prod_{i=1}^{j} w_t(x_i)\right] = \mathbf{E}_{\{x_1,\ldots,x_j\}}\left[\prod_{i=1}^{N_t} w_0\left(\xi_t^i\right)\right], \qquad \text{Lebesgue-a.e.}
$$

A first application: large-scale behaviour

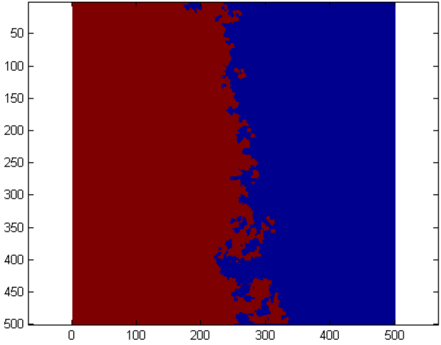
Initial configuration:



Simulations by H. Saadi. Fixed radius, $u \equiv 1$.

A first application: large-scale behaviour

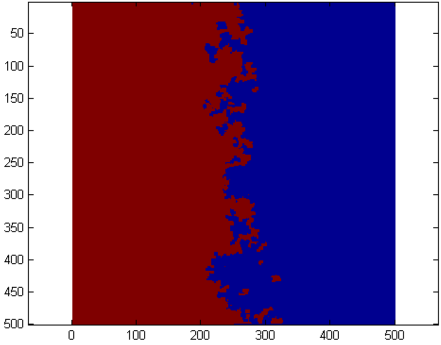
After 2.10^6 events:



Simulations by H. Saadi. Fixed radius, $u \equiv 1$.

A first application: large-scale behaviour

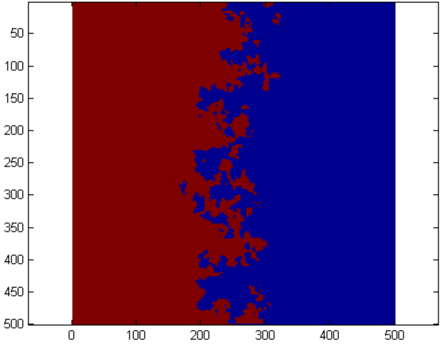
After 3.10^6 events:



Simulations by H. Saadi. Fixed radius, $u \equiv 1$.

A first application: large-scale behaviour

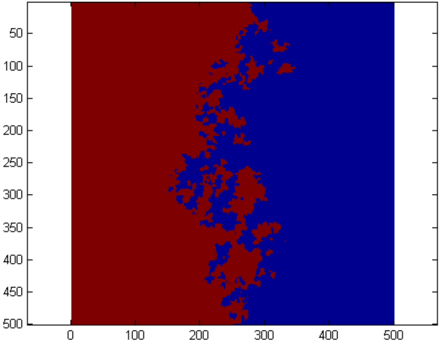
After 4.10^6 events:



Simulations by H. Saadi. Fixed radius, $u \equiv 1$.

A first application: large-scale behaviour

After 5.10^6 events:



Simulations by H. Saadi. Fixed radius, $u \equiv 1$.

Large-scale evolution (with N. Berestycki & A.E.)

Geographical space: $\mathbb{R}^d$, **Type space:** $\{0, 1\}$

- **Case 1: Fixed radii**

  We fix $R > 0$ and $u \in (0, 1]$. All events have radius R and impact u.

  $\hookrightarrow$ Most natural first case...

  $\hookrightarrow$ Asymptotic behaviour equivalent to that of the nearest-neighbour stepping-stone model.

Large-scale evolution (with N. Berestycki & A.E.)

Geographical space: $\mathbb{R}^d$, **Type space:** $\{0,1\}$

- **Case 1: Fixed radii**

  We fix $R > 0$ and $u \in (0,1]$. All events have radius R and impact u.

  $\hookrightarrow$ Most natural first case...

  $\hookrightarrow$ Asymptotic behaviour equivalent to that of the nearest-neighbour stepping-stone model.

- **Case 2: Radii with an α-stable distribution**

  We fix an impact $u \in (0,1]$, $\alpha \in (1,2)$ and take as a measure on radii

  $$\mu(dr) = \frac{\mathbf{1}_{\{r>1\}}}{r^{d+1+\alpha}}\, dr.$$

  $\hookrightarrow$ Allows very large but very rare events.

  $\hookrightarrow$ Rescaled ancestral lineages are well-understood.

Zoom-out

- Case 1: Fixed radius and impact
- Case 2: Fixed impact and intensity of radii $r^{-(d+\alpha+1)}\,dr$

Set $\alpha = 2$ in case 1, and for all $n \geq 1$,

$$w_t^n(x) := w_{nt}(n^{1/\alpha} x).$$

Zoom-out

- Case 1: Fixed radius and impact
- Case 2: Fixed impact and intensity of radii $r^{-(d+\alpha+1)}\,dr$

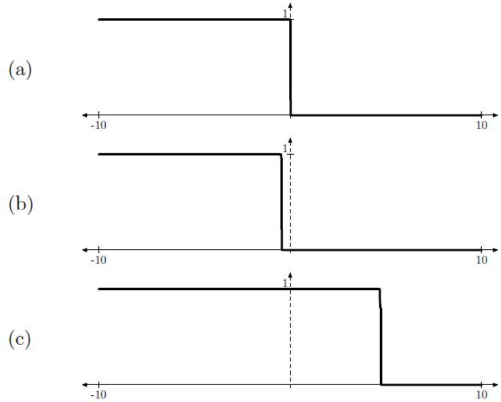
Set $\alpha = 2$ in case 1, and for all $n \geq 1$,

$$w_t^n(x) := w_{nt}(n^{1/\alpha}x).$$

Initial condition: $w_0(x) = \mathbf{1}_H(x)$, where $H = \{x_{(1)} \leq 0\}$.

Questions: What does w_t^n look like when n is large? Width of the interface? Pattern of genetic diversity? Roughness of the interface?

Answer for fixed radius, $d = 1$



$u = 0.8$, $r = 0.033$ and $n = 10^3$. Initial condition, after 10^5 events, after 10^7 events. (Simulations by J. Kelleher)

That is...

Theorem 1 [Berestycki, Etheridge & V. (2012)]

- There exists a measure valued process $(M_t^{(2)},\ t \geq 0)$ such that

$$M^n \xrightarrow{(fdd'\,s)} M^{(2)}, \qquad \text{as } n \to \infty.$$

- Moreover, one can find $\tilde{\sigma}^2 > 0$ such that, if X denotes BM and

$$p_t^{(2)}(x) := \mathbb{P}\big[X_{u\tilde{\sigma}^2 t} \in H\big], \text{ then}$$

That is...

Theorem 1 [Berestycki, Etheridge & V. (2012)]

- There exists a measure valued process $(M_t^{(2)},\ t \geq 0)$ such that

$$M^n \overset{(fdd's)}{\longrightarrow} M^{(2)}, \qquad \text{as } n \to \infty.$$

- Moreover, one can find $\tilde{\sigma}^2 > 0$ such that, if X denotes BM and

$$p_t^{(2)}(x) := \mathbb{P}\big[X_{u\tilde{\sigma}^2 t} \in H\big], \text{ then}$$

$\hookrightarrow$ If $d = 1$: for every $t > 0$, $w_t^{(2)}$ is a random field of correlated Bernoulli r.v.'s with

$$\mathbb{E}\big[w_t^{(2)}(x)\big] = p_t^{(2)}(x).$$

That is...

Theorem 1 [Berestycki, Etheridge & V. (2012)]

- There exists a measure valued process $(M_t^{(2)},\, t \geq 0)$ such that

$$M^n \xrightarrow{(fdd's)} M^{(2)}, \qquad \text{as } n \to \infty.$$

- Moreover, one can find $\tilde{\sigma}^2 > 0$ such that, if X denotes BM and

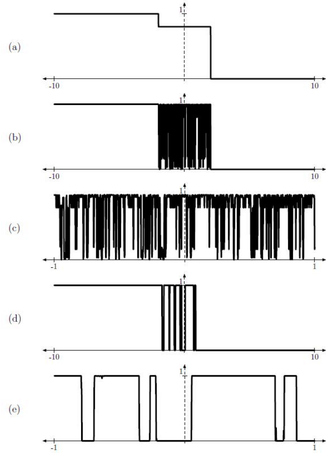
$$p_t^{(2)}(x) := \mathbb{P}\big[X_{u\tilde{\sigma}^2 t} \in H\big], \text{ then}$$

$\hookrightarrow$ If $d = 1$: for every $t > 0$, $w_t^{(2)}$ is a random field of correlated Bernoulli r.v.'s with

$$\mathbb{E}\big[w_t^{(2)}(x)\big] = p_t^{(2)}(x).$$

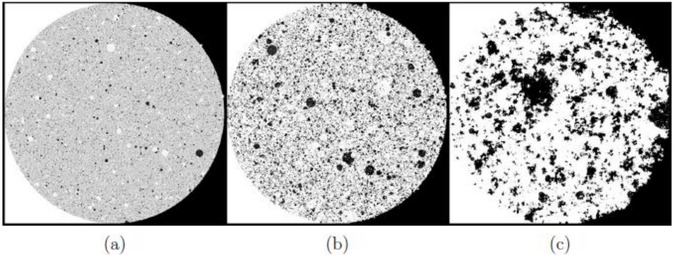
$\hookrightarrow$ If $d \geq 2$: for every $t \geq 0$, $w_t^{(2)}(x) = p_t^{(2)}(x)$ Lebesgue-a.e.

Case of stable radii, $d = 1$



$u = 0.8$, $\alpha = 1.3$ and $n = 10^4$.
(a) Initial condition, (b-c) after 100 events, (d-e) after 10^6 events.

Case of stable radii, $d = 2$



$u = 0.8$, $\alpha = 1.3$ and $n = 10^3$. After 10^5, 10^6 and 10^7 events.

Asymptotic behaviour in the presence of large events

Theorem 2 [Berestycki, Etheridge & V. (2012)]

- There exists a measure valued process $(M_t^{(\alpha)}, t \geq 0)$ such that

$$M^n \xrightarrow{(fdd's)} M^{(\alpha)}, \qquad \text{as } n \to \infty.$$

- Moreover, there exists a symmetric α-stable process $X^{(\alpha)}$ such that, if

$$p_t^{(\alpha)}(x) := \mathbb{P}\big[X_{ut}^{(\alpha)} \in H\big]$$

then *in any dimension*, for every $t > 0$, $w_t^{(\alpha)}$ is a random field of correlated Bernoulli r.v.'s with

$$\mathbb{E}\big[w_t^{(\alpha)}(x)\big] = p_t^{(\alpha)}(x).$$

Conclusions

- **No coexistence of types** unless $d \geq 2$ and reproduction is 'purely local'.
- The impact **u appears only in the limiting speed** of evolution (same pattern of allele frequencies for all $u \in (0,1]$);
- The correlations between local frequencies are given by the genealogical process. **Correlation length:**
  - $\sqrt{n}$ when only small events,
  - $n^{1/\alpha}$ when mixture of events.
- Since $n^{1/\alpha} \gg \sqrt{n}$, this neutral model can explain the **correlation lengths much larger than expected** in certain pops.

  $\Rightarrow$ Large but rare extinction/recolonization can have a significant impact on the genetic diversity of a population.

Idea of the proof

- By duality, for every $j \geq 1$ and $\psi \in C_c((\mathbb{R}^d)^j)$,

$$\mathbb{E}_{w_0^n}\Big[\int_{(\mathbb{R}^d)^j} \psi(x_1,\ldots,x_j)\Big\{\prod_{i=1}^{j} w_t^n(x_i)\Big\} dx_1\cdots dx_j\Big]$$
$$= \int_{(\mathbb{R}^d)^j} \psi(x_1,\ldots,x_j)\mathbf{E}_{\{x_1,\ldots,x_j\}}\Big[\prod_{i=1}^{N_t} w_0^n(\xi_t^{n,i})\Big] dx_1\cdots dx_j,$$

  where

$$w_0^n = \mathbf{1}_H \quad \text{and} \quad \xi_t^{n,i} = n^{-1/\alpha}\,\xi_{nt}^i.$$

- These test functions characterize the law of each M_t.
  $\Rightarrow$ Understanding the limit of ξ^n gives the limit of w^n.

Idea of the proof

- By duality, for every $j \geq 1$ and $\psi \in C_c((\mathbb{R}^d)^j)$,

$$\mathbb{E}_{w_0^n}\left[\int_{(\mathbb{R}^d)^j} \psi(x_1,\dots,x_j)\left\{\prod_{i=1}^{j} w_t^n(x_i)\right\} dx_1\cdots dx_j\right]$$

$$= \int_{(\mathbb{R}^d)^j} \psi(x_1,\dots,x_j)\mathbf{E}_{\{x_1,\dots,x_j\}}\left[\prod_{i=1}^{N_t} w_0^n(\xi_t^{n,i})\right] dx_1\cdots dx_j,$$

  where

$$w_0^n = \mathbf{1}_H \quad \text{and} \quad \xi_t^{n,i} = n^{-1/\alpha}\,\xi_{nt}^i.$$

- These test functions characterize the law of each M_t.
  $\Rightarrow$ Understanding the limit of ξ^n gives the limit of w^n.

- **Correlations:**

$$\mathbb{E}_{w_0^n}\left[\prod_{i=1}^{j} w_t^n(x_i)\right] = \mathbf{P}_{\{x_1,\dots,x_j\}}\left[\xi_t^{n,i} \in H,\ \forall i \in \{1,\dots,N_t^n\}\right].$$

Genealogies in the limit

Under local events:

1 lineage After rescaling, an ancestral line jumps at rate $\mathcal{O}(n)$ at distance $\mathcal{O}(1/\sqrt{n})$

$\Rightarrow$ A single lineage converges to Brownian motion, with speed $\sigma^2 = u\tilde{\sigma}^2$.

More lineages Two lineages

$\hookrightarrow$ move independently when at distance $> 2R/\sqrt{n}$,

$\hookrightarrow$ may coalesce only when at distance $\leq 2R/\sqrt{n}$.

$\Rightarrow$ The ancestral process converges to a system of independent Brownian motions which coalesce upon meeting.

Genealogies in the limit

Under local events:

1 lineage After rescaling, an ancestral line jumps at rate $\mathcal{O}(n)$ at distance $\mathcal{O}(1/\sqrt{n})$

$\Rightarrow$ A single lineage converges to Brownian motion, with speed $\sigma^2 = u\tilde{\sigma}^2$.

More lineages Two lineages

$\hookrightarrow$ move independently when at distance $> 2R/\sqrt{n}$,
$\hookrightarrow$ may coalesce only when at distance $\leq 2R/\sqrt{n}$.

$\Rightarrow$ The ancestral process converges to a system of independent Brownian motions which coalesce upon meeting.

Under mixed events: The ancestral process converges to a system of coalescing symmetric α-stable processes. A finite sample reaches its MRCA in finite time a.s.

Back to original scales

- Under the assumption of local reproduction, the evolution over large scales depends only on σ^2.
- Cannot be the case when we consider small to intermediate geogr.- and time-scales (coalescence is not instantaneous, e.g.).

  $\Rightarrow$ **Other quantities summarizing the local evolution?**
- Even when large but rare bottlenecks occur, they will not be seen over sufficiently small scales (genealogies ‘resolved’ in a few hundred generations only).

The Wright-Malécot formula

As in the stepping-stone model, let us set

$$F_\mu(|x-y|) := \mathbf{E}_{\{x,y\}}\left[e^{-2\mu T_c}\right].$$

The Wright-Malécot formula

As in the stepping-stone model, let us set

$$F_\mu(|x-y|) := \mathbf{E}_{\{x,y\}}\left[e^{-2\mu T_c}\right].$$

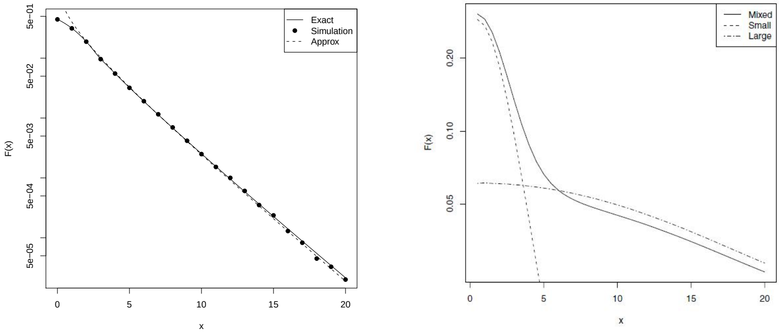
When reproduction is purely local and $\mu \ll 1$, F_μ is well-approximated by the *Wright-Malécot formula*:

$$F_\mu(|x-y|) \approx \frac{K_0(|x-y|/\ell_\mu)}{\mathcal{N} + \log(\ell_\mu/\kappa)}, \qquad |x-y| > \kappa$$

where

- $\ell_\mu = \sigma/\sqrt{2\mu} \gg 1$ is a *characteristic length*;
- κ is a *local scale* given by the precise local dynamics;
- $\mathcal{N}$ measures the *number of potential parents* of an individual ($\propto 1/u$ here).

In pictures



Fit between F_μ (plain lines) and the Wright-Malécot formula (dashed lines).

Left: local rep. only; Right: 2 types of events. (Figures by J. Kelleher)

Frequency-based inference

- σ^2, $\mathcal{N}$ and κ summarize the local evolution of genetic diversities.
- Assume mutation occurs at rate $\mu \ll 1$ and maintains an average heterozygosity H_μ over some intermediate spatial scale.
- Using the duality formula, we obtain

$$\frac{\mathrm{Cov}(\rho(x), \rho(y))}{H_\mu} \approx \mathbf{E}\left[e^{-2\mu T_c}\right] \approx \frac{K_0(|x-y|/\ell_\mu)}{\mathcal{N} + \log(\ell_\mu/\kappa)}.$$

Frequency-based inference

- σ^2, $\mathcal{N}$ and κ summarize the local evolution of genetic diversities.
- Assume mutation occurs at rate $\mu \ll 1$ and maintains an average heterozygosity H_μ over some intermediate spatial scale.
- Using the duality formula, we obtain

$$\frac{\mathrm{Cov}(\rho(x), \rho(y))}{H_\mu} \approx \mathbf{E}\left[e^{-2\mu T_c}\right] \approx \frac{K_0(|x-y|/\ell_\mu)}{\mathcal{N} + \log(\ell_\mu/\kappa)}.$$

- **A basis for inference:** Call $\overline{H}$ the average heterozygosity in a sample taken from nearby sites $x_1, \ldots, x_n$. If $x_i \neq x_j$,

$$\frac{\mathrm{Cov}(\rho(x_i), \rho(x_j))}{\overline{H}} \approx \frac{K_0(|x_i - x_j|/\ell_\mu)}{\mathcal{N}} \tag{1}$$

- Assuming the frequencies are Gaussian fluct. around their mean, (1) yields a maximum likelihood scheme [Barton et al, 2012].

Correlations across loci

Correlations across loci

- **Question :** We understand well the genealogies at 1 locus, what about more than 1? A whole genome?
- **Main characteristic:** Two recombinants may coalesce again quickly, for ex. due to the next event which overlaps them.

  ⇒ Creates potentially strong correlations between the allele frequencies at neighbouring loci.

Correlations across loci

- **Question :** We understand well the genealogies at 1 locus, what about more than 1? A whole genome?
- **Main characteristic:** Two recombinants may coalesce again quickly, for ex. due to the next event which overlaps them.

  $\Rightarrow$ Creates potentially strong correlations between the allele frequencies at neighbouring loci.
- **Sub-questions:**
  - $\hookrightarrow$ Are there regimes of parameters for which decorrelation between the ancestral lineages of an individual at two (or more) loci can occur ? What are the local mechanisms maintaining some correlations?
  - $\hookrightarrow$ Influence of the presence of large extinction/recolonization events?
  - $\hookrightarrow$ Difference with the pattern left behind by a selective sweep? by recurrent global bottlenecks?

On the scale of the whole population

Geographical space: $\mathbb{R}^2$, **Type space:** $K_1 \times K_2$ (2 loci)

Again, 2 types of events:

Small ev. Each site is hit at rate $\mathcal{O}(1)$ by an event of size $\mathcal{O}(1)$.

$\hookrightarrow$ A random number of parents is chosen;

$\hookrightarrow$ A fraction u_s of the local population is killed.

$\hookrightarrow$ A fraction r_n of the offspring are *recombinants* (i.e., inherit their types k_1, k_2 from different parents)

On the scale of the whole population

Geographical space: $\mathbb{R}^2$, **Type space:** $K_1 \times K_2$ (2 loci)

Again, 2 types of events:

Small ev. Each site is hit at rate $\mathcal{O}(1)$ by an event of size $\mathcal{O}(1)$.

$\hookrightarrow$ A random number of parents is chosen;

$\hookrightarrow$ A fraction u_s of the local population is killed.

$\hookrightarrow$ A fraction r_n of the offspring are *recombinants* (i.e., inherit their types k_1, k_2 from different parents)

Large ev. Each site is hit at rate ϕ_n^{-1} by an event of size $\mathcal{O}(n^\alpha)$, where $\alpha > 0$. A fraction u_B of the local pop. is replaced, and we assume no recombination for simplicity.

Regime $1 \ll \phi_n \ll n^{2\alpha}$ as $n \to \infty$, and $(r_n)_{n\geq 1}$ is nonincreasing.

On the scale of the whole population

Geographical space: $\mathbb{R}^2$, **Type space:** $K_1 \times K_2$ (2 loci)

Again, 2 types of events:

Small ev. Each site is hit at rate $\mathcal{O}(1)$ by an event of size $\mathcal{O}(1)$.

- $\hookrightarrow$ A random number of parents is chosen;
- $\hookrightarrow$ A fraction u_s of the local population is killed.
- $\hookrightarrow$ A fraction r_n of the offspring are *recombinants* (i.e., inherit their types k_1, k_2 from different parents)

Large ev. Each site is hit at rate ϕ_n^{-1} by an event of size $\mathcal{O}(n^\alpha)$, where $\alpha > 0$. A fraction u_B of the local pop. is replaced, and we assume no recombination for simplicity.

Regime $1 \ll \phi_n \ll n^{2\alpha}$ as $n \to \infty$, and $(r_n)_{n\geq 1}$ is nonincreasing.

Sample 2 individuals at distance $x_n \gg n^\alpha$.

$\Rightarrow$ Joint distribution of the coal. time at the two loci, as $n \to \infty$?

Patterns of correlations across loci

Theorem [Etheridge & V. (2012)]

- If we sample 2 individuals at distance $x_n \gg n^\alpha$, the genealogy at each locus is Kingman's coalescent when considered on the timescale
$$\phi_n\, n^{2(t-\alpha)},\ t > \alpha.$$

Patterns of correlations across loci

Theorem [Etheridge & V. (2012)]

- If we sample 2 individuals at distance $x_n \gg n^\alpha$, the genealogy at each locus is Kingman's coalescent when considered on the timescale

$$\phi_n \, n^{2(t-\alpha)}, \; t > \alpha.$$

- In addition, there exists a critical distance

$$D_n^* \approx n^\alpha \sqrt{1 + \frac{\log \phi_n}{r_n \phi_n}}$$

  such that when n is large,

  $\hookrightarrow$ If $x_n \gg D_n^*$, the ancestries at the two loci are independent,

  $\hookrightarrow$ If $x_n \ll D_n^*$, there is a *decorrelation threshold* before which the genealogies are completely correlated, and after which they become approximately independent.

Conclusions

- $$(\phi_n/n^{2\alpha})\ n^{2t} \ll n^{2t},$$

  ⇒ Large events generate a faster coalescence, and so (again) much larger correlation lengths between allele frequencies.

- The second result gives us the sampling distance at which we should expect to see a decorrelation between the variations in allele freq. at the two loci, with or without large events.

  ⇒ Comparison with the effect of sweeps possible.

- **But** sampling distances must be very large. Locally, the probability of decorrelation is very small.

  ⇒ Consider instead many loci (or a long continuous genome).

Length of regions identical in state

- Assume only **local reproduction** (but robust to rare large events);
- **Many loci**, with recombination rate r between 2 neighbours;

Length of regions identical in state

- Assume only **local reproduction** (but robust to rare large events);
- **Many loci**, with recombination rate r between 2 neighbours;

- Sample 2 individuals at **small/medium distance** δ.
- Consider the regions of the genetic map where the two individuals are **identical in state**, in particular the large blocks generated by *early* coalescence.

Length of regions identical in state

- Assume only **local reproduction** (but robust to rare large events);
- **Many loci**, with recombination rate *r* between 2 neighbours;

- Sample 2 individuals at **small/medium distance** δ.
- Consider the regions of the genetic map where the two individuals are **identical in state**, in particular the large blocks generated by *early* coalescence.

- **Early coalescence** means on a timescale of order $(\delta/\sigma)^2$, where σ^2 is the variance of the motion of a lineage.

  $\Rightarrow$ for some $\beta > 0$, set

  $$\mu(\beta, \delta) = \frac{\sigma^2}{2\beta\delta^2} \qquad \text{and} \qquad T_\mu \sim \mathrm{Exp}(2\mu).$$

  A coalescence at locus *j* is *early* if $T_c^j \leq T_\mu$.

An approximation

Theorem [Barton et al. (2012)]

Let X be the length of a given region of identity in state generated by an early coal., when the two indv. are sampled at distance δ.

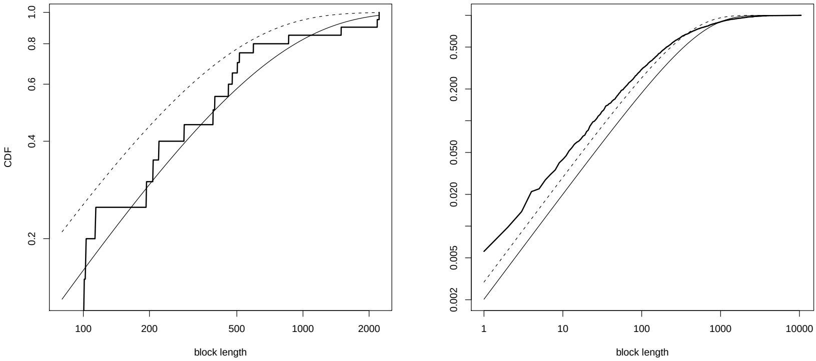
Then X follows approximately a geometric distribution with parameter $\gamma(\delta)$ given by

$$\gamma(\delta) = \frac{r_{\text{eff}}}{r_{\text{eff}} + \mu} \left(1 - \frac{K_0(1/\sqrt{\beta})}{\mathcal{N} + \log(\sqrt{\beta}\,\delta/\kappa)} \right),$$

where

- κ and $\mathcal{N}$ come from the Wright-Malécot approx.,
- $r_{\text{eff}} = r\,\psi(\delta)$ is an *effective recombination rate*,
- $\psi(\delta)$ is the **escape probability** of two recombinant lineages.

Simulations (by J. Kelleher)



CDF of long conserved blocks, (*left*) from a single sim. and (*right*) from 200 sim. $R = 1$, $u = 0.75$, $r = 10^{-5}$, $\delta = 10$ and 50k loci.

Heavy solid line: empirical; Dashed line: $\mathrm{Geom}(\gamma(\delta))$; Solid line: $\mathrm{Geom}(\hat{p})$.

Still a lot of work...

- The parameter $\gamma(\delta)$ depends 'only' on σ^2, $\mathcal{N}$ and κ.
  $\Rightarrow$ Another route to **inference**?

Still a lot of work...

- The parameter $\gamma(\delta)$ depends 'only' on σ^2, $\mathcal{N}$ and κ.

  $\Rightarrow$ Another route to **inference**?

- **Several problems:**

  $\hookrightarrow$ The empirical CDF overestimates the probability of large regions (genealogies are embedded in the same *pedigree*).

  $\hookrightarrow$ Not easy to relate regions of identity in state between the 2 genomes, and regions of early coalescence. In particular, which β should we take ?

Further questions

Natural selection

We bias the choice of the parent, by giving a weight $1 + s$ to type 1 individuals, and weight 1 to type 0 indv.
$\Rightarrow$ **Dual available**, but branches as well (potential selection events).

Natural selection

We bias the choice of the parent, by giving a weight $1 + s$ to type 1 individuals, and weight 1 to type 0 indv.
$\Rightarrow$ **Dual available**, but branches as well (potential selection events).

- **Large neighbourhood size:** when the impact u_n and the selection strength s_n tend to 0 appropriately,

  $\hookrightarrow$ **In $1d$ and with only local rep.**, the frequency of type 1 individuals (suitably rescaled) converges to the solution to

  $$dw = \frac{1}{2}\Delta w\, dt + \tilde{s}w(1-w)\, dt + \sqrt{\frac{1}{N_e}\, w(1-w)}\, B(dt, dx),$$

  where $B(dt, dx)$ is a space-time white noise.

  $\hookrightarrow$ **In higher dim.**, no noise in the limit.

  $\hookrightarrow$ Equivalent results when **large-scale bottlenecks** occur, and only the motion is affected (still a local selection pressure and local coalescence).

  (Work in progress with A. Etheridge and F. Yu.)

Natural selection

We bias the choice of the parent, by giving a weight $1+s$ to type 1 individuals, and weight 1 to type 0 indv.
$\Rightarrow$ **Dual available**, but branches as well (potential selection events).

- **Large neighbourhood size:** when the impact u_n and the selection strength s_n tend to 0 appropriately,
  - $\hookrightarrow$ **In $1d$ and with only local rep.**, the frequency of type 1 individuals (suitably rescaled) converges to the solution to

$$dw = \frac{1}{2}\,\Delta w\,dt + \tilde{s}w(1-w)\,dt + \sqrt{\frac{1}{N_e}\,w(1-w)}\,B(dt,dx),$$

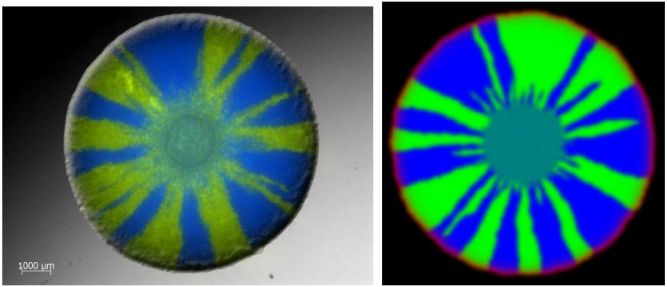
    where $B(dt,dx)$ is a space-time white noise.
  - $\hookrightarrow$ **In higher dim.**, no noise in the limit.
  - $\hookrightarrow$ Equivalent results when **large-scale bottlenecks** occur, and only the motion is affected (still a local selection pressure and local coalescence).

  (Work in progress with A. Etheridge and F. Yu.)

- **Small neighbourhood size:** The pattern produced is very different (cf. Nick's presentation).

Range expansion

Extreme case of selection: only type 1's reproduce.



Expanding population of Pseudomonas aeruginosa (courtesy of Kevin Foster), and a simulation of the modified SLFV, by J. Kelleher.

(Work in progress with A. Etheridge and J. Kelleher)

Thank you!