Ancestries of a Recombining Diploid Population:

towards a theoretical union of behavioural ecology and population genetics

Raazesh Sainudiin[†]

[†] Department of Mathematics, Uppsala University, Uppsala, Sweden
Joint work with: Bhalchandra Thatte (Brasil) and Amandine Veber (France)

September 12 2017

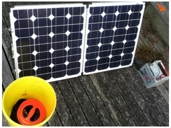
CoSy Lunch Seminar, Centre for Interdisciplinary Mathematics, Uppsala University, Uppsala, Sweden

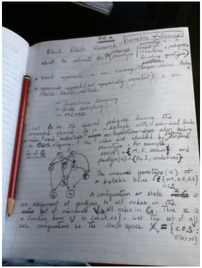
Insights from the field

Ancestries of Recombining Diploid Populations

Insights from field work in November 2011

Insights from field work in November 2011





POA — Genetics + Science

Black Robin Research

want to estimate (θ). P(observed phenotype | parameters in model of inheritance, pedigree, founding genotypes)

- Exact approach is via Cannings–Thompson–Skolnick Peeling
- approximate approach (but asymptotically consistent) is via Monte Carlo methods.
  - Importance Sampling
  - Gibbs Sampling
  - MCMC

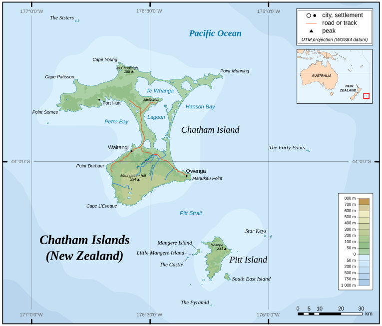
Let G be the observed pedigree during the managed period. G is a digraph with individual birds as nodes and ~~marriage edges~~ ~~or copulation edges~~ edges leading to each offspring. The nodes are identified by observed sex and phenotype. For example, $sex(v_i) \in \{M, F, \text{unkown}\}$ and $phenotype(v_i) \in \{0, 1, \text{undefined}\}$

Eg of G

The unobserved $genotype(v_i)$ at a biallelic locus $\in \{aa, aA, AA\} =: S$

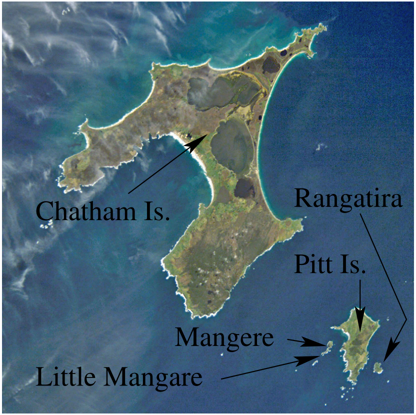
A configuration or state x ~~of G~~ is an assignment of genotypes to all nodes in the ~~vertex~~ set of individuals V i.e. all nodes in G. Thus x is a function from V to $\{aa, aA, AA\}$. Let the set of all such configurations be the state space $\mathbb{X}_\pi = \{x \in S^V : \pi(x) > 0\}$

Map of Antarctica / Chatham Islands (New Zealand)



- The Chatham Islands lie 800 kms east of Christchurch.
- Two main islands – Chatham Island and Pitt Island

Chatham Islands (New Zealand)

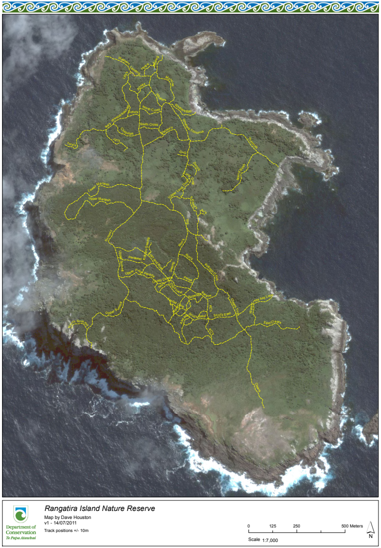


- Many smaller islands – including Little Mangere, Mangere and Rangatira
- The islands have only emerged above sea level in the last 4 million years and are part of the Chathams rise connected below ocean to NZ.
- The climate is cool and wet with salt-laden winds all year round.

Chatham Islands (New Zealand)

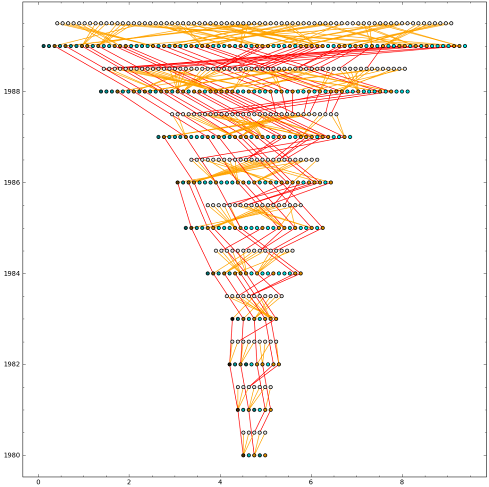
Rangatira Island

- is the third largest island in the Chatham Islands
- covers an area of 218 hectares (539 acres)
- is a gazetted nature reserve since 1953
- is now home to many endemic species
- is home to black robins – saved from near extinction by team led by Don Merton in the 1980s



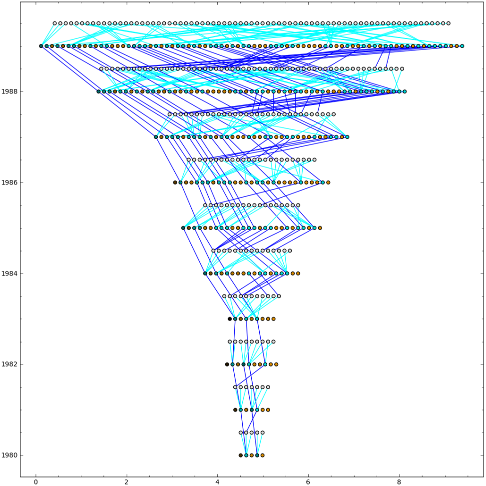
A Chatham island Black Robin in Rangatira today

Temporal Pedigrees with Overlapping Generations



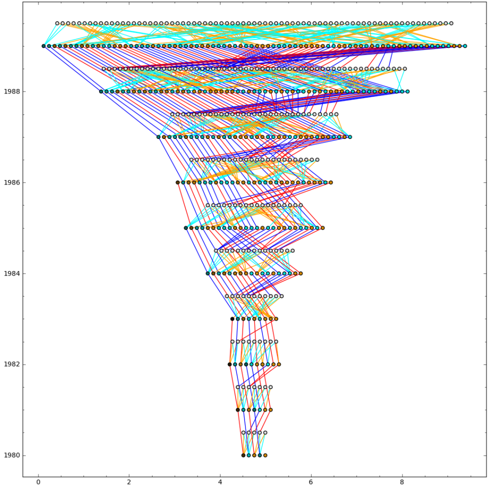
Female Population Tree during Field Conservation in 1980s

Temporal Pedigrees with Overlapping Generations



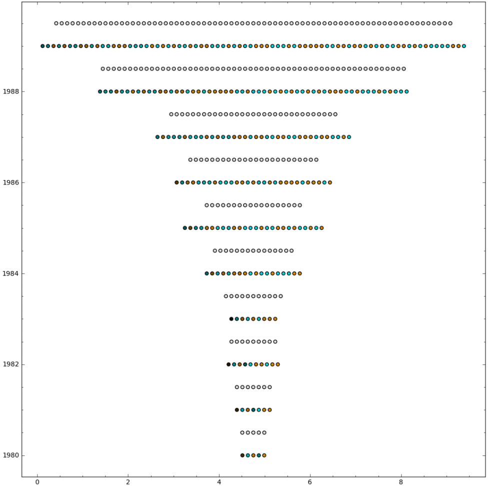
Male Population Tree during Field Conservation in 1980s

Temporal Pedigrees with Overlapping Generations



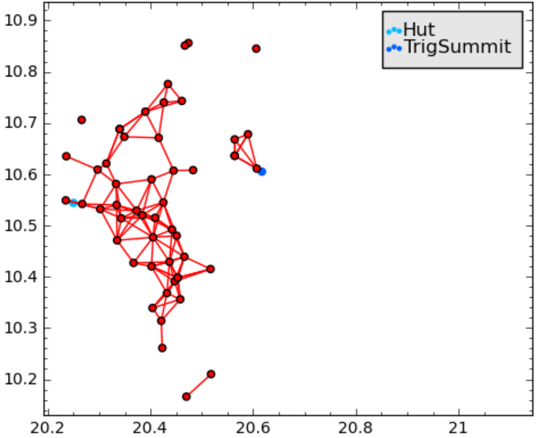
Population Pedigree during Field Conservation in 1980s

Temporal Pedigrees with Overlapping Generations



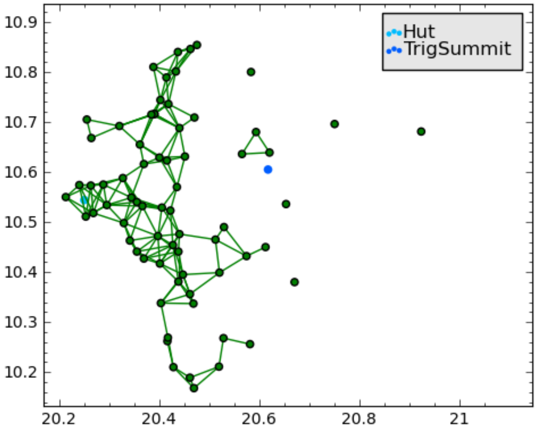
Coalescent Sub-Pedigree of the 1989 population $\Longrightarrow$ Part ARDP

Male Territory Graphs by year



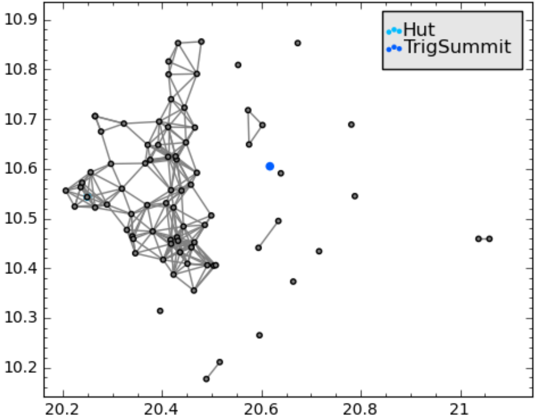
2009 (eye-balled by a ball graph)

Male Territory Graphs by year



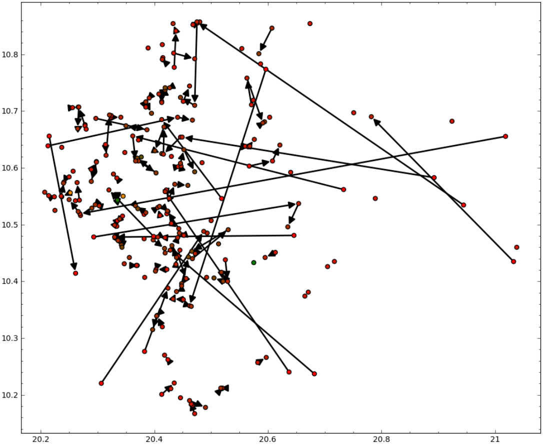
2010 (eye-balled by a ball graph)

Male Territory Graphs by year

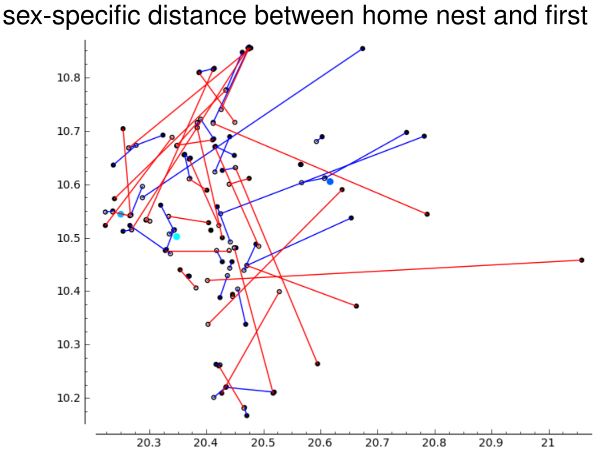


2011 (eye-balled by a ball graph)

Male Territorial Dynamics – yearly nest site relocation

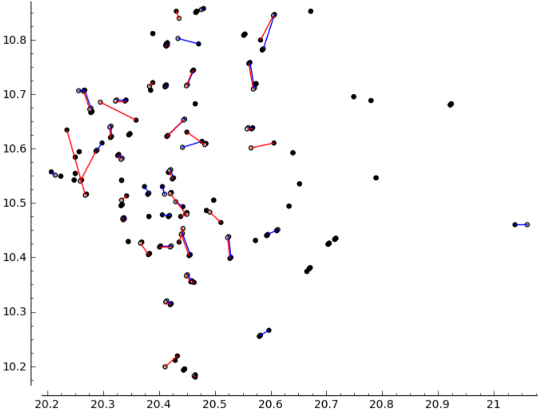


Male / Female yearly nest to nest distance

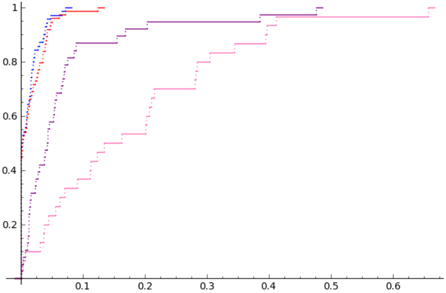


Male / Female yearly nest to nest distance

sex-specific distance between nest at age ≥ 1 and next nest



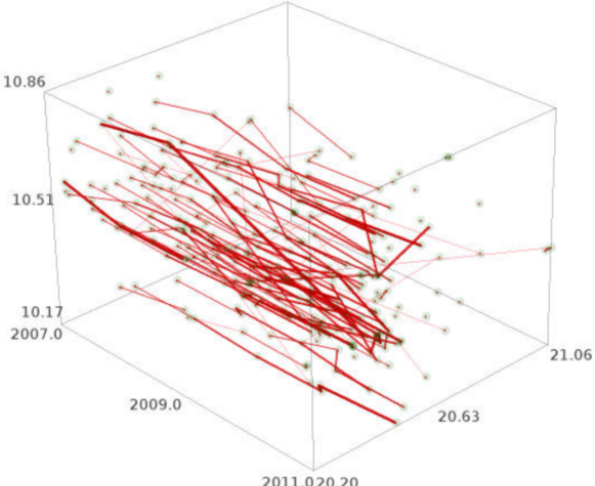
Male / Female yearly nest to nest distance



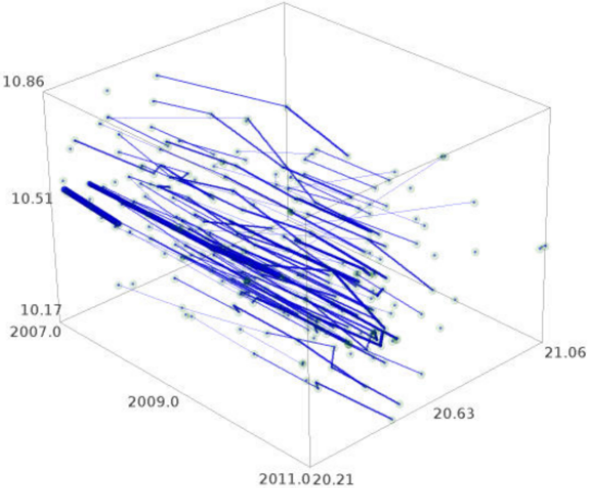
- Permutation Test to Reject H_0 : male dist = female dist (10^5 MC samples)
- Strongly reject H_0 for home to first nest (pv = 0.00025)
- Fail to reject H_0 for nest at age ≥ 1 to next nest (pv = 0.14)

male / female distance from home nest to first nest
male / female distance from nest at age $\geq$ 1 to next nest

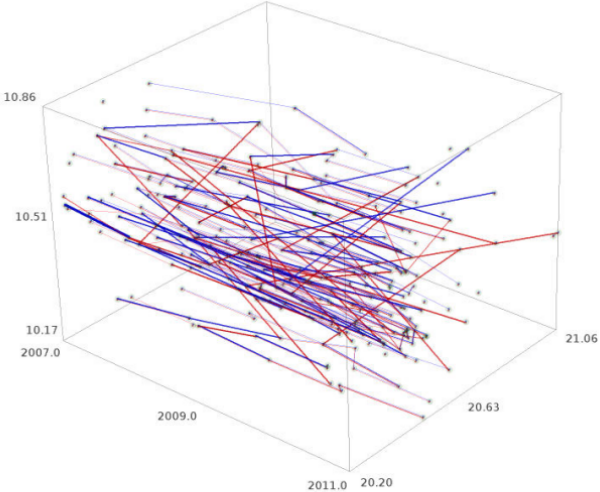
Need Models of Spatio-temporal Behavioural Population Pedigrees



Need Models of Spatio-temporal Behavioural Population Pedigrees



Need Models of Spatio-temporal Behavioural Population Pedigrees



parametric families of behavioural pedigree processes

- *a simulation-intensive study:* "Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent". (Wakeley et al. Genetics, 2012).

parametric families of behavioural pedigree processes

- *a simulation-intensive study:* “Gene genealogies within a fixed pedigree, and the robustness of Kingman’s coalescent”. (Wakeley et al. Genetics, 2012).
- from similar simulations: However there are *gender-specific behavioural effects on the structure of the spatio-temporal pedigree through which genes flow under natural selection* – think of **Florida Scrub Jay** with cooperative breeding strategies, polygamous terriotorial feral cat populations, Mujahidiin heroms, fruit flies, **some nematode species**...

parametric families of behavioural pedigree processes

- *a simulation-intensive study:* “Gene genealogies within a fixed pedigree, and the robustness of Kingman’s coalescent”. (Wakeley et al. Genetics, 2012).
- from similar simulations: However there are *gender-specific behavioural effects on the structure of the spatio-temporal pedigree through which genes flow under natural selection* – think of **Florida Scrub Jay** with cooperative breeding strategies, polygamous terriotorial feral cat populations, Mujahidiin heroms, fruit flies, **some nematode species**...
- monogamy/andry-permiscuity parameter via a sex-specific generator of *historical mate/re-mate choice (HMC)*.

parametric families of behavioural pedigree processes

- *a simulation-intensive study:* “Gene genealogies within a fixed pedigree, and the robustness of Kingman’s coalescent”. (Wakeley et al. Genetics, 2012).
- from similar simulations: However there are *gender-specific behavioural effects on the structure of the spatio-temporal pedigree through which genes flow under natural selection* – think of **Florida Scrub Jay** with cooperative breeding strategies, polygamous terriotorial feral cat populations, Mujahidiin heroms, fruit flies, **some nematode species**...
- monogamy/andry-permiscuity parameter via a sex-specific generator of *historical mate/re-mate choice (HMC)*.
- monogamy/andry-polygamy/andry-permiscuity parameter with sex-specific HMC generator $\oplus$ sex-ratios.

parametric families of behavioural pedigree processes

- *a simulation-intensive study:* “Gene genealogies within a fixed pedigree, and the robustness of Kingman’s coalescent”. (Wakeley et al. Genetics, 2012).
- from similar simulations: However there are *gender-specific behavioural effects on the structure of the spatio-temporal pedigree through which genes flow under natural selection* – think of **Florida Scrub Jay** with cooperative breeding strategies, polygamous terriotorial feral cat populations, Mujahidiin heroms, fruit flies, **some nematode species**...
- monogamy/andry-permiscuity parameter via a sex-specific generator of *historical mate/re-mate choice (HMC)*.
- monogamy/andry-polygamy/andry-permiscuity parameter with sex-specific HMC generator $\oplus$ sex-ratios.
- patri/matri-locality parameters with sex-specific spatial dispersal kernels.

parametric families of behavioural pedigree processes

- *a simulation-intensive study:* “Gene genealogies within a fixed pedigree, and the robustness of Kingman’s coalescent”. (Wakeley et al. Genetics, 2012).
- from similar simulations: However there are *gender-specific behavioural effects on the structure of the spatio-temporal pedigree through which genes flow under natural selection* – think of **Florida Scrub Jay** with cooperative breeding strategies, polygamous terriotorial feral cat populations, Mujahidiin heroms, fruit flies, **some nematode species**...
- monogamy/andry-permiscuity parameter via a sex-specific generator of *historical mate/re-mate choice (HMC)*.
- monogamy/andry-polygamy/andry-permiscuity parameter with sex-specific HMC generator $\oplus$ sex-ratios.
- patri/matri-locality parameters with sex-specific spatial dispersal kernels.
- simple parametrizations *seem* identifiable (from simulations at the finest pedigree resolution).

parametric families of behavioural pedigree processes

- *a simulation-intensive study:* "Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent". (Wakeley et al. Genetics, 2012).
- from similar simulations: However there are *gender-specific behavioural effects on the structure of the spatio-temporal pedigree through which genes flow under natural selection* – think of **Florida Scrub Jay** with cooperative breeding strategies, polygamous terriotorial feral cat populations, Mujahidiin heroms, fruit flies, **some nematode species**...
- monogamy/andry-permiscuity parameter via a sex-specific generator of *historical mate/re-mate choice (HMC)*.
- monogamy/andry-polygamy/andry-permiscuity parameter with sex-specific HMC generator $\oplus$ sex-ratios.
- patri/matri-locality parameters with sex-specific spatial dispersal kernels.
- simple parametrizations *seem* identifiable (from simulations at the finest pedigree resolution).
- Other problems with Robustness of Kingman's Coalescent: (i) $n \to N$, (ii) complex trait mapping may also need "the-5-dot-diploid-model" – two autosomes, two sex chromosomes and the cytosol tracked in a topologically explicit manner on the same $(\Omega, \mathcal{F}, \mathcal{P})$

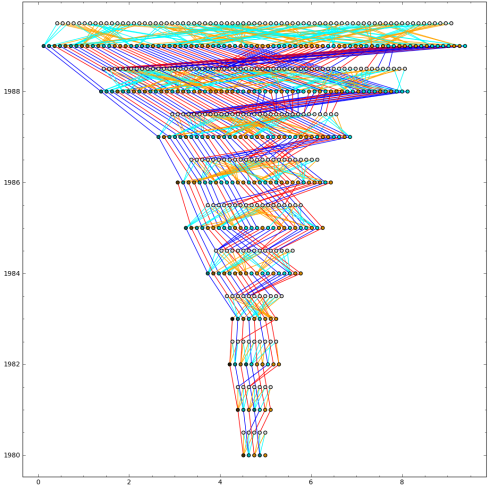
parametric families of behavioural pedigree processes

- *a simulation-intensive study:* "Gene genealogies within a fixed pedigree, and the robustness of Kingman's coalescent". (Wakeley et al. Genetics, 2012).
- from similar simulations: However there are *gender-specific behavioural effects on the structure of the spatio-temporal pedigree through which genes flow under natural selection* – think of **Florida Scrub Jay** with cooperative breeding strategies, polygamous terriotorial feral cat populations, Mujahidiin heroms, fruit flies, **some nematode species**...
- monogamy/andry-permiscuity parameter via a sex-specific generator of *historical mate/re-mate choice (HMC)*.
- monogamy/andry-polygamy/andry-permiscuity parameter with sex-specific HMC generator $\oplus$ sex-ratios.
- patri/matri-locality parameters with sex-specific spatial dispersal kernels.
- simple parametrizations *seem* identifiable (from simulations at the finest pedigree resolution).
- Other problems with Robustness of Kingman's Coalescent: (i) $n \rightarrow N$, (ii) complex trait mapping may also need "the-5-dot-diploid-model" – two autosomes, two sex chromosomes and the cytosol tracked in a topologically explicit manner on the same $(\Omega, \mathcal{F}, \mathcal{P})$
- **CORE MODEL** *(Dominating discrete skeletal measure)*: Coalescent with Recombination inside "A Well-stirred Urn of Self-Fertile Hermaphrodites"

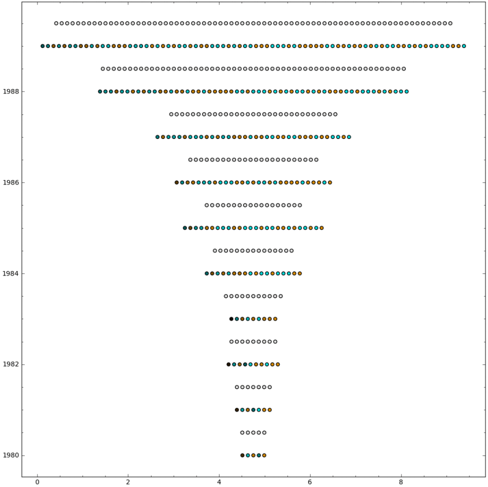
Part ARDP

Ancestries of Recombining Diploid Populations

Raazesh Sainudiin, Bhalchandra Thatte, and Amandine Véber,
Journal of Mathematical Biology, January 2016, Volume 72, Issue
1-2, pp 363–408.



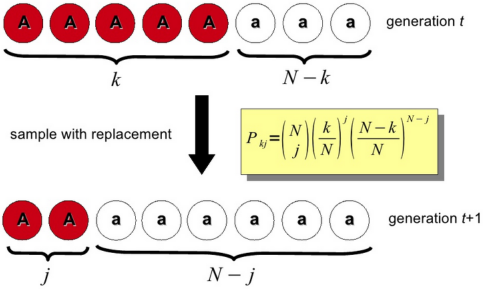
Motivation: Diploid Population Pedigree (forwards in time)



Coalescent Sub-Pedigree (backwards in time) “inheritance topology of diploid zygotes?”

Three Notions of Molecular continuums within W-F Population Pedigrees

One *r* to rule them all!

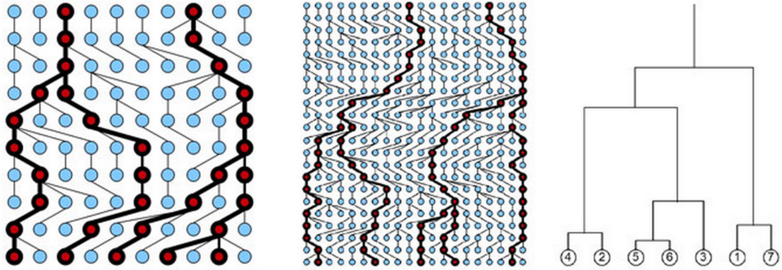


Wright-Fisher Model (forwards in time)

where $r = 0$ is the probability of recombination per locus per meiotic generation

Three Notions of Molecular continuums within W-F Population Pedigrees

One *r* to rule them all!

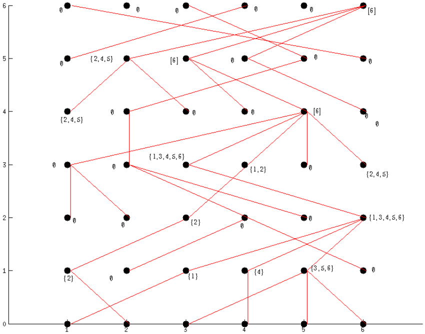


Wright-Fisher Genealogy with Large N $\Longrightarrow$ Kingman's Coalescent Tree for small *n* (backwards in time)

where *r* is the probability of recombination per locus per meiotic generation ($r = 0$ for Kingman's model)

Three Notions of Molecular continuums within W-F Population Pedigrees

One *r* to rule them all!

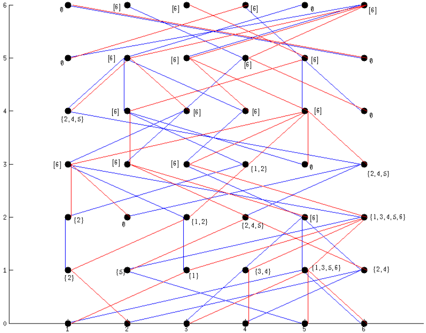


(1) Kingman's Discrete Genealogy at a non-recombining locus with $r = 0$ (forwards in time)

Observe: Time to most recent ancestor (TMRCA) = 4 (can you locate the MRCA?)

Three Notions of Molecular continuums within W-F Population Pedigrees

One *r* to rule them all!



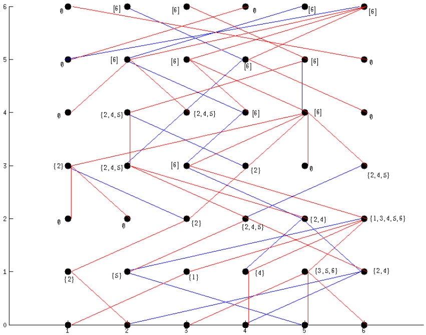
(2) Chang's Discrete Pedigree as the Genealogy of an always-recombining locus with $r = 1$ (forwards in time)

Observe: Time to most recent ancestor (TMRCA) = 2 (can you locate the MRCA? how about other CAs?)

Observe: Time to be or not to be (TTBONTB) = 5 (can you say why?)

Three Notions of Molecular continuums within W-F Population Pedigrees

One *r* to rule them all!



(3) Hudson-Griffiths Discrete Ancestral Recombination Graph at a recombining locus with $0 < r < 1$ (fwds in time)
SIMPLEST UNIFICATION PROBLEM: What is TMRCA and TTBONTB for any r?

(NOT for $rN =: \rho > 0$, where $r \to 0$ and $N \to \infty$, the Griffiths-Husdon limit)

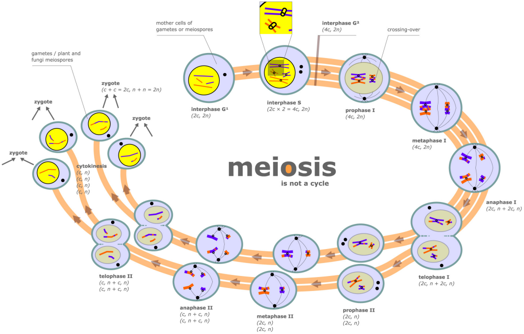
Unifying these three notions of molecular ancestry

Consider the **ancestral graphs** of these three genealogical processes for a given sample taken today (backwards in time).

- Uni-parental Ancestry of sample : Kingman's Discrete Genealogy at a non-recombining locus (1982)
- Bi-parental Ancestry of sample : Chang's Discrete Pedigree (1999)
- Uni/Bi-parental Ancestry of sample: Hudson-Griffiths Discrete Ancestral Recombination Graph (1983-1989)

Unifying these three notions of molecular ancestry

consistent with the zygotic topology in sync with meiosis:



This unification was an open problem since 1983

This topologically consistent notion of diploid ancestry (RS, Thatte & Veber, 2015) in a common probability space is a pre-requisite for more realistic models of population pedigree processes, including:

- non-overlapping generations

This unification was an open problem since 1983

This topologically consistent notion of diploid ancestry (RS, Thatte & Veber, 2015) in a common probability space is a pre-requisite for more realistic models of population pedigree processes, including:

- non-overlapping generations
- dioecious (non-hermaphroditic) diploidy

This unification was an open problem since 1983

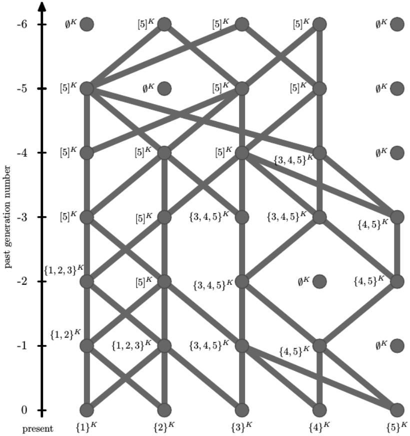
This topologically consistent notion of diploid ancestry (RS, Thatte & Veber, 2015) in a common probability space is a pre-requisite for more realistic models of population pedigree processes, including:

- non-overlapping generations
- dioecious (non-hermaphroditic) diploidy
- gender-specific behavioural effects on spatio-temporal pedigree through which genes flow under natural selection...

This unification was an open problem since 1983

This topologically consistent notion of diploid ancestry (RS, Thatte & Veber, 2015) in a common probability space is a pre-requisite for more realistic models of population pedigree processes, including:

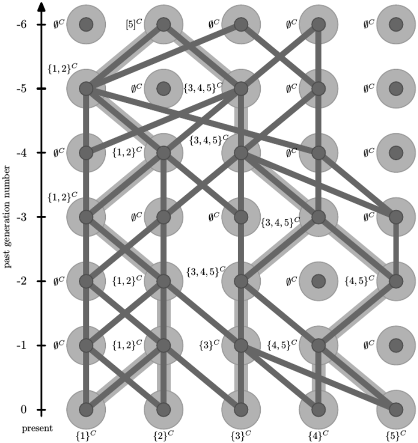
- non-overlapping generations
- dioecious (non-hermaphroditic) diploidy
- gender-specific behavioural effects on spatio-temporal pedigree through which genes flow under natural selection...
- etc. (think of Florida Scrub Jay, feral cat populations, heroms of Mujahidiin warlords post Soviet withdrawal in 1980s)



karyotic ancestral graph of a bi-parental Wright-Fisher population with five eukaryotic individuals

Can you find the MRC-karyotic-A (MRCkA)? TMRCkA? TT-karyotically-BONTB (TTkBONTB)?

Chang's Main results say that TMRCkA and TTkBONTB $\sim O(\log_2(n))$, where $n =$ population size



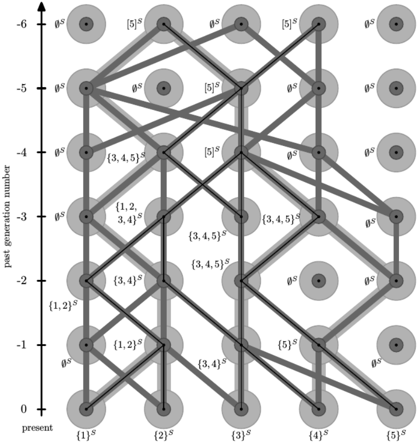
Zygotic pedigree containing the cytoplasmic tree (Kingman's discrete coalecent) and karyotic ancestral graph of a

bi-parental Wright-Fisher population

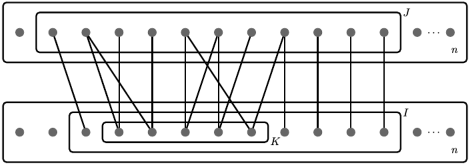
Can you find the MRC-cytoplasmic-A (MRCcA)? TMRCcA?

Kingman's Main results say that TMRCcA and TTcBONTB $\sim$ $O(n)$, where $n =$ population size

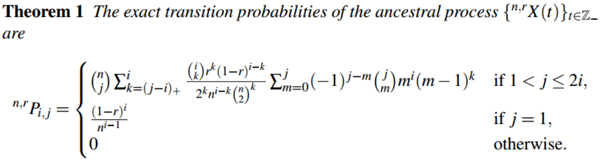
so: $O(\log_2(n))$ versus $O(n)$ is a huge difference!



Putting it all together: Zygotic, cytoplasmic, karyotic and sub-karyotic ancestral graphs (discrete ARGs of Hudson and Griffiths) of a bi-parental Wright-Fisher population with five eukaryotic diploid individuals



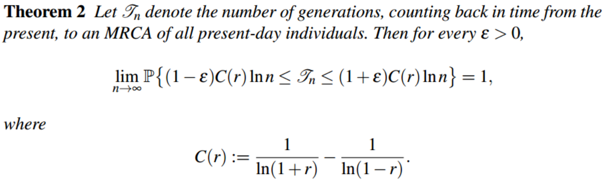
Combinatorial Structure Diagram



Theorem 1 *The exact transition probabilities of the ancestral process* $\{{}^{n,r}X(t)\}_{t\in\mathbb{Z}_-}$ *are*

$$ {}^{n,r}P_{i,j} = \begin{cases} \binom{n}{j}\sum_{k=(j-i)_+}^{i} \frac{\binom{i}{k} r^k (1-r)^{i-k}}{2^k n^{i-k} \binom{n}{2}^k} \sum_{m=0}^{j} (-1)^{j-m} \binom{j}{m} m^i (m-1)^k & \text{if } 1 < j \leq 2i, \\ \frac{(1-r)^i}{n^{i-1}} & \text{if } j = 1, \\ 0 & \text{otherwise.} \end{cases} $$

Kingman-Chang-Hudson-Griffiths Unified Ancestral Chain



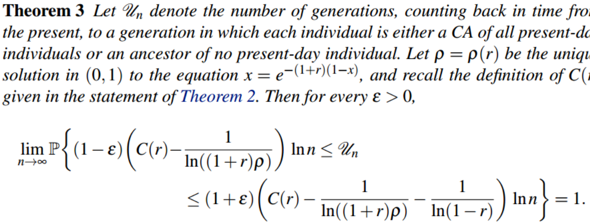
Theorem 2 *Let $\mathscr{T}_n$ denote the number of generations, counting back in time from the present, to an MRCA of all present-day individuals. Then for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{(1-\varepsilon)C(r)\ln n \leq \mathscr{T}_n \leq (1+\varepsilon)C(r)\ln n\right\} = 1,$$

where

$$C(r) := \frac{1}{\ln(1+r)} - \frac{1}{\ln(1-r)}.$$

Time to MRCA for a possibly recombining locus with intra-locus recombination probability per non-overlapping meiotic generation in a population of *n* diploid individuals.



Theorem 3 *Let $\mathscr{U}_n$ denote the number of generations, counting back in time from the present, to a generation in which each individual is either a CA of all present-day individuals or an ancestor of no present-day individual. Let $\rho = \rho(r)$ be the unique solution in $(0,1)$ to the equation $x = e^{-(1+r)(1-x)}$, and recall the definition of $C(r)$ given in the statement of Theorem 2. Then for every $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left\{(1-\varepsilon)\left(C(r) - \frac{1}{\ln((1+r)\rho)}\right)\ln n \le \mathscr{U}_n \le (1+\varepsilon)\left(C(r) - \frac{1}{\ln((1+r)\rho)} - \frac{1}{\ln(1-r)}\right)\ln n\right\} = 1.$$

Time to Be or Not to Be for a possibly recombining locus with intra-locus recombination probability per non-overlapping meiotic generation in a population of *n* diploid individuals.

Proof ideas

Sketches of the Proofs of Thms 1 & 2 is a 50 minute chalk talk.

Proof ideas

Sketches of the Proofs of Thms 1 & 2 is a 50 minute chalk talk.

- For Thm 1: Just count the integers in the numerator and denominator of the rational probabilities
    - counting Wright-Fisher constrained bipartite graphs from the combinatorial diagram
- For Theorems 2 and 3:
    - carefully partitioning the time-axis into six distinct n-asymptotic regimes

$$\left[0, \oplus \frac{2\log\log n}{\log(1+r)}, \oplus \frac{\log n}{\log(1+r)}, \oplus \log\log n, \oplus \log\log n, \oplus - \frac{\log n}{\log(1-r)}, \oplus \frac{-2\log\log n}{\log(1-r)}\right]$$

and applying the right tools from below:

- $\approxeq$ the $\#$(descendants / non-descendants of individual I_i), binomial RVs, by a GW branching process with Poisson($1 + / - r$) offspring distribution (start/end)
- Bernstein $\neq$'s for the Binomial RVs away from start/end
- standard branching process results from Athreya & Ney

Conclusions

- Just the first steps here –
  Kingman-Chang-Hudson/Griffiths r-parametric family.

Conclusions

- Just the first steps here –
  Kingman-Chang-Hudson/Griffiths r-parametric family.
- Empirical Population Pedigree Processes should be informed by animal behavior (field insights $\mapsto$ math-stat)

Conclusions

- Just the first steps here –
  Kingman-Chang-Hudson/Griffiths r-parametric family.
- Empirical Population Pedigree Processes should be informed by animal behavior (field insights $\mapsto$ math-stat)
- Behavioral data for animal populations is far from idealized limiting probability models:
  random mating, non-overlapping generations,...

Conclusions

- Just the first steps here –
  Kingman-Chang-Hudson/Griffiths r-parametric family.
- Empirical Population Pedigree Processes should be informed by animal behavior (field insights $\mapsto$ math-stat)
- Behavioral data for animal populations is far from idealized limiting probability models:
  random mating, non-overlapping generations,...
- Need “combinatorially rich enough” statistical models

Conclusions

- Just the first steps here –
  Kingman-Chang-Hudson/Griffiths r-parametric family.
- Empirical Population Pedigree Processes should be informed by animal behavior (field insights $\mapsto$ math-stat)
- Behavioral data for animal populations is far from idealized limiting probability models:
  random mating, non-overlapping generations,...
- Need “combinatorially rich enough” statistical models
- Such models may pave a way for the **fusion** of behavioural ecology and evolutionary genetics
  A paradigm for "Population Structure From Behavior"

Conclusions

- Just the first steps here –
  Kingman-Chang-Hudson/Griffiths *r*-parametric family.
- Empirical Population Pedigree Processes should be informed by animal behavior (field insights $\mapsto$ math-stat)
- Behavioral data for animal populations is far from idealized limiting probability models:
  random mating, non-overlapping generations,...
- Need “combinatorially rich enough” statistical models
- Such models may pave a way for the **fusion** of behavioural ecology and evolutionary genetics
  A paradigm for "Population Structure From Behavior"
- PLAN: continue to work with FSJ @CornellLab’O (2016)

Thank you!

Funds for Theory:

- School of Maths & Stats (solar panels for remote computing in 2011; Sabbatical grant 2014),
- Research Chaire in 2013: Vieola Environment - French Natural History Museum - Ecole Polytechnique.