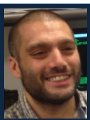
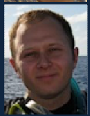
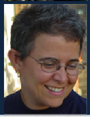

Edward A. Salinas (1)


John B. Horne (2,3)


Rita Castilho (3)

COMET

COalescence with Multiple mergers Employing Thermodynamic integration

CCMAR
Centre of Marine Sciences

(1) Independent Programmer/Researcher, Dallas, TX, USA
(2) Dalhousie University, Halifax, Nova Scotia, Canada
(3) CCMAR, Univ. do Algarve, Faro, Portugal

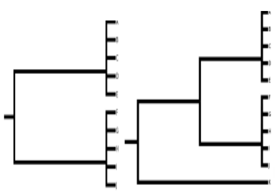
Calls and Motivations for Multiple Mergers Genealogies

The Kingman coalescent has been a one-size-fits-all model for population genealogies. Others however have recently been developed offering trees with multiple mergers (nodes with three or more children) which differ from the Kingman coalescent, viz., overlapping generations, reproductive variance, and no limit on the offspring an individual can contribute to the population size[1]. Such models are particularly relevant for marine organisms with high larval mortality and extreme reproductive variance and any species with overlapping generations. Indeed, there have been calls for such models [2].

1. Tellier A., Lemaire C., Coalescence 2.0: a multiple branching of recent theoretical developments and their applications , Molecular Ecology (2014) Volume 23, Issue 11, pp 2637–2652
2. Marra A., et al. (2015) Population Genetic History of Aristeus antennatus (Crustacea: Decapoda) in the Western and Central Mediterranean Sea. PLoS ONE 10(3): e0117272. doi:10.1371/journal.pone.0117272

Algorithms for Multiple Mergers Genealogies

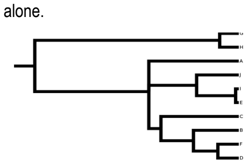
To explore multiple mergers genealogies and models, we selected two models from [3], each expressed with an algorithm. **Algorithm #1** provides for generating genealogies with both **multiple mergers** and **generational overlap**. Its parameter *Y* is used to model a "*degree of simultaneity*" for multiple merger events. For example, if *Y*=2, then half of the lineages might coalesce to a single parent and the remaining lineages in another parent, both coalescent events happening at the same time. Its parameter *Φ* may be used as a "*generational overlap factor*" yielding a percentage of lineages whose coalescence is delayed.



Algorithm 1.
(1) Start with *K* = *n* (*K* is the number of the ancestors of the sample as we follow them back in time).
(2) Sample (*Y*, *Φ*) from a joint distribution
(3) Generate the waiting time to the first SRE (looking bakwards in time): it is an exponential random variable with mean one.
(4) Sample (K_0, K_1,.....K_Y) from *Mn*(K, p_0, p_1,.....p_Y), where $\sum_{i=0}^{Y} p_0 = 1 - \phi$ and $p_i = \phi/Y, i = 1, \ldots, Y$. (Here *Mn* (·) denotes the multinomial distribution).
(5) If $K_i \neq 0$, *i* = 1, ….,*Y* , then randomly choose K_i individuals from *K* and merge them into a common ancestor.
(6) Reset the number of ancestors to $K = K_0 + \sum_{i=0}^{Y} 1\{k_i \neq 0\}$, where $1\{x\}$ is one if the *x* is true and zero, otherwise. Stop if *K* is equal to one, otherwise return to Step 2.

Trees generated with Algorithm #1, Y=2, *Φ*=0, B. Y=2, *Φ*=0.9 ; Algorithm #1 Made with FigTree [4]

Algorithm #2[3] provides Algorithm #1 but also it contains the Kingman Coalescent in a hybrid fashion by alternating between Kingman style coalescent events and multiple mergers events as in Algorithm #1 by a distribution of **"successful reproduction events"** (**SREs**). Such a combination may improve Algorithm #2's model fit relative to either the Kingman or Algorithm #1 alone.



A tree generated by[4] and algorithm #2 with Y=1, *Φ*=.71

Algorithm 2.
(1) Start with *K* = *n*.
(2) Sample (*Y*, *Φ*) from a joint distribution.
(3) Generate the waiting time to the first SRE (looking bakwards in time): it is an exponential random variable with mean one.
(4) Up to time *t*, run the usual binary, the Kingman coalescent with rate *K*(*K*-1)/*c*, while there are *K* lineages. Form pairwise common ancestors, as coalescent events occur, and decrement K by one, at each event. Exit the algorithm if *K* = 1 before time *t*.
(5) Using the value of K(>1) from Step 4, sample (K_0, K_1,.....K_Y) from Mn(*K*, p_0, p_1,.....p_Y), where $p_0 = 1 - \phi$ and $p_i = \phi/Y, i = 1, \ldots, Y$.
(6) If $K_i \neq 0$, *i*=1, ….,*Y* then randomly choose K_i individuals from *K* without replacement and coalesce them together.
(7) Reset to $K = K_0 + \sum_{i=0}^{Y} 1\{k_i \neq 0\}$, where $1\{x\}$ is one if the *x* is true and zero, otherwise. Stop if *K* is equal to one, otherwise return to Step 2.

3. Ori Sargsyan, John Wakeley, A coalescent process with simultaneous multiple mergers for Approximating the gene genealogies of many marine organisms, Theoretical Population Biology 74 (2008) 104–11
4. FigTree, available from http://tree.bio.ed.ac.uk/software/figtree/

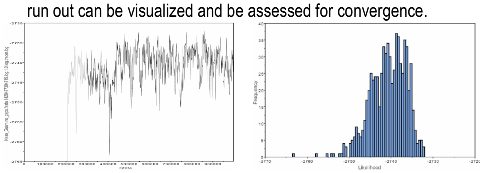
Introducing COMET : COalescence with Multiple mergers Employing Thermodynamic integration

MCMC (or Markov Chain Monte Carlo) is a well-known method in bioinformatics for sampling from **posterior distributions of model parameters** given data [5-8]. To do so, we have implemented an open-source MCMC program which incorporates i) the Kingman Coalescent, ii) Algorithm #1 and iii) Algorithm #2. This way, our program samples from the space of multiple mergers genealogies. Since our code is accompanied by routines to perform **thermodynamic integration** for calculating **marginal probabilities** for Bayes Factors and model selection, we call our program **COMET : COalescence with Multiple mergers Employing Thermodynamic integration.** COMET takes as input i) a **FASTA** file of a multiple sequence alignment and ii) a configuration file for setting up MCMC runs (iterations, priors, etc.). It is written in the Java programming language and is available for download on bitbucket : http://www.bitbucket.org/eddieasalinas/comet

5. Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. Biometrika, 57, 97–109
6. Ziheng Y., Molecular Evolution, A Statistical Approach, Oxford Univ. Press, 2014
7. Wakely, J. Coalescent Theory: An Introduction, Roberts and Co. Publishers, 2009
8. Felsenstein, J., Inferring Phylogenies, Sinauer Press, 2003

Trace files for Tracer and Visual Inspection and Assessment of both Mixing and Convergence

COMET generate log files containing sampled values from posterior distributions, for example: trees, their likelihoods, model parameters (e.g. *Y*, *Φ*). An included utility script "tracer_prep.pl" takes log files generated by COMET and makes them suitable for the **Tracer** [9] program by removing data columns not suitable for tracer (such as trees, timestamps, etc.). Once fed into tracer, the run out can be visualized and be assessed for convergence.

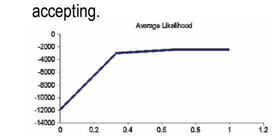


A likelihood trace from tracer (left) and its posterior frequency histrogram (right) from Tracer [9] .

9. Rambaut A, Suchard MA, Xie D & Drummond AJ (2014) Tracer v1.6, Available from http://beast.bio.ed.ac.uk/Tracer

Thermodynamically Integrated Marginal Probabilities for Bayes Factors and Model Selection

Thermodynamic Integration (TI) is for computing **marginal probabilities P(D|M)** [10]; the ratio of two such values defines a **Bayes Factor** often used for model selection. COMET implements TI with various *β* settings from 0,...,1 for MCMC chains with heated acceptance. Hotter chains are more accepting.



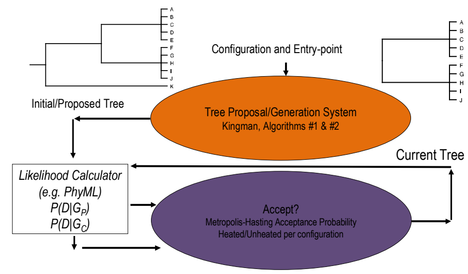
According to previous research [10], the harmonic mean (HM) method for marginal likelihoods is more variable than TI.

10. Peter Beerli, Michal Palczewski, Unified Framework to Evaluate Panmixia and Migration Direction among Multiple Sampling Locations, Genetics 185: 313–326 (May2010)

A graph showing TI with avg. log likelihoods from MCMC chains with heats in the range 0,..,1

Calculating Likelihoods of Genealogies

MCMC requires computation of **likelihoods**. In COMET's case it is the probability of the data given a genealogy *P(D|G)* [10]. COMET incorporates the **HKY85** [11,12] model for computing likelihoods with PAL [13]. Incorporating **PhyML** [14,15] for the **likelihood calculations** [10] however has led to a dramatic improvement for performance. Work is in progress to incorporate **BEAGLE** [16]. The schematic below indicates how the likelihood and tree code are incorporated into the overall MCMC scheme.

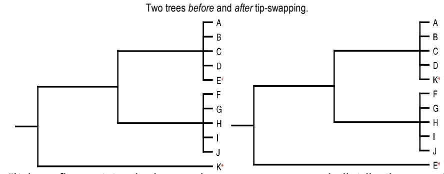


10. Felsenstein J., Evolutionary Trees from DNA Sequences : A Maximum Likelihood Approach, Journal of Molecular Evolution 1981, Volume 17, Issue 6, pp 368-376
11 Hasegawa M, Kishino H, Yano T., Dating of the human-ape splitting by a molecular clock of mitochondrial DNA., J Mol Evol. 1985;22(2):160-74
12. Yang, Z., Computational Molecular Evolution, 2006, Oxford Univ. Press, 2006
13. Alexei Drummond, Korbinian Strimmer, PAL: an object-oriented programming library for molecular evolution and phylogenetics, 17 (7): 662-663
14. Guindon S, Gascuel O., A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood, Syst Biol. 2003 Oct;52(5):696-704
15. Guindon S., et al.., New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0, Systematic Biology, 59(3):307-21, 2010
16. Ayres, et. al., BEAGLE: An Application Programming Interface and High-Performance Computing Library for Statistical Phylogenetics, Syst Biol. 2012 Jan; 61(1): 170–173

Challenges and Opportunities in Assessing and Bringing About Convergence : Proposals of Proposals

Numerous MCMC runs were executed, **ESS (effective sample size)** values[6,17] observed, and log files combined with **"logcombiner"**[17] as we tested COMET. To improve mixing and convergence, we have introduced a proposal method "tip swapping" where a pair of random leaves (sequences) are swapped in the tree to create a proposed tree from the current one. Implementation and incorporation into COMET seems to have dramatically improved mixing and convergence.

Two trees *before* and *after* tip-swapping.



"It is a fine art to design and compose proposal distributions and the operators that implement them. The efficiency of MCMC algorithms crucially depends on a good mix of operators forming the proposal distribution." (p. 16) [18] We continue to experiment and explore with this and other schemes and strategies.

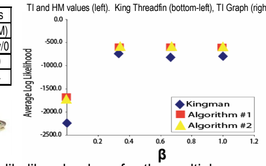
17. Drummond AJ, Suchard MA, Xie D & Rambaut A (2012) Bayesian phylogenetics with BEAUti and the BEAST 1.7 Molecular Biology And Evolution 29: 1969-1973
18. Bayesian evolutionary analysis with BEAST Alexei J. Drummond and Remco R. Bouckaert Cambridge University Press, 2015

An example workflow : Computation of a Bayes Factor for Model Selection Favoring a Multiple Mergers Model

To explore COMET's capabilities we analyzed with COMET *v0.0.72* a dataset of 10 DNA sequences from the mitochondrial control region of *Polydactylus macrochir (King threadfin)*[19]. Ambiguous bases and gaps were removed. Four runs at *β* values *0, 1/3, 2/3* and *3/3* were done the Kingman coalescent, Algorithm #1, and Algorithm #2. The priors were set as follows : *Y* uniform over *{1,2,3}* ; *Φ* uniform over *[0.5,1.0]*. *c* was set at *4* during the Kingman and Algorithm #2 runs. *5,000,000* iterations and burn-in of *4,000,000* were used per run and each batch of four runs took about six to eight hours to complete.

TI and HM values (left). King Threadfin (bottom-left), TI Graph (right)

	Marginal Likelihoods	
Model	TI P(D|M)	HM P(D|M)
Kingman	-805.13	underflow/0
Algorithm #1	-618.30	-585.29
Algorithm #2	-625.62	-587.84



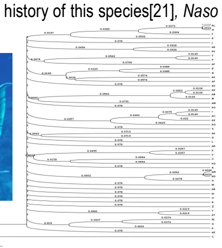
The TI an HM marginal likelihood values for the multiple mergers models (-618 and -626) vs. the Kingman model (-805) yield Bayes Factors ratios favoring multiple mergers models.

19. John B. Horne, Paolo Momigliano, David J. Welch, Stephen J. Newman, Lynne van Herwerden, Searching for common threads in threadfins: phylogeography of Australian polynemids in space and time, Mar Ecol Prog Ser 449: 263–276, 2012 ; Genbank Accessions : NT0199K,NT0200K,NT0201K,NT0202K,NT0203K,NT0211K,NT0213K,NT0216K,NT0219K,NT0220K

Genealogy Visualization and Interpretation

There are multiple mergers towards the root of the tree [20], and not the base, and these speak of a sudden population expansion that occurred long ago in the demographic history of this species[21], *Naso unicornis*, the Unicorn Fish.

Photo of N. Unicornis from John B. Horne, PhD



20. H. Harpending and A. Rogers, Genetic Perspectives on Human Origins and Differentiation, Annu. Rev. Genomics Hum. Genet. 2000. 1:361–85
21. Horne, et al., Limited eco. pop. Conn. suggests low demands on self-recruitment in a tropical inshore marine fish (Eleutheronema tetradactylum: Polynemidae), Mol. Eco (2011) 20, 2291–2306

Next Steps and Future Directions

Current and future work to enhance COMET includes improving the likelihood calculations with BEAGLE[16] and/or newer software such as the **PLL : Phylogenetic Likelihood Library** [22]. We plan to experiment with other proposal schemes such as **"neighborhood rearrangements"**[23] as well. We also plan to explore with other datasets such as : **shrimp (*Aristeus antennatus*), barnacles (*Balanus glandula*) , or maybe even humans (*Homo sapiens sapiens*)**.

22. T. Flouri,1, et al., The Phylogenetic Likelihood Library, Syst Biol. 2015 Mar; 64(2): 356–362
23. Kuhner, M. K., Yamato, J. and Felsenstein, J. (1995) Estimating effective population size and mutation rate from sequence data using Metropolis-Hastings sampling. Genetics, 140, 1421–1430.