

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23

On the independent loci assumption in phylogenomic studies

W. Bryan Jennings

Departamento de Vertebrados, Museu Nacional, Universidade Federal do Rio de Janeiro, Rio de Janeiro, RJ, 20940-040, Brazil.

Email address: wbjenn@gmail.com

24 **Abstract**

25 Studies using multi-locus coalescent methods to infer species trees or historical demographic
26 parameters usually require the assumption that the gene tree for each locus (or SNP) is
27 genealogically independent from the gene trees of other sampled loci. In practice, however,
28 researchers have used two different criteria to delimit independent loci in phylogenomic studies.
29 The first criterion, which directly addresses the condition of genealogical independence of
30 sampled loci, considers the long-term effects of homologous recombination and effective
31 population size on linkage between two loci. In contrast, the second criterion, which only
32 considers the single-generation effects of recombination in the meioses of individuals, identifies
33 sampled loci as being independent of each other if they undergo Mendelian independent
34 assortment. Methods that use these criteria to estimate the number of independent loci per
35 genome as well as intra-chromosomal “distance thresholds” that can be used to delimit
36 independent loci in phylogenomic datasets are reviewed. To compare the efficacy of each
37 criterion, they are applied to two species (an invertebrate and vertebrate) for which relevant
38 genetic and genomic data are available. Although the independent assortment criterion is
39 relatively easy to apply, the results of this study show that it is overly conservative and therefore
40 its use would unfairly restrict the sizes of phylogenomic datasets. It is therefore recommended
41 that researchers only refer to *genealogically* independent loci when discussing the independent
42 loci assumption in phylogenomics and avoid using terms that may conflate this assumption with
43 independent assortment. Moreover, whenever feasible, researchers should use methods for
44 delimiting putatively independent loci that take into account both homologous recombination
45 and effective population size (i.e., long-term effective recombination).

46

47 **Introduction**

48 A key assumption of phylogenomic studies that use multi-locus coalescent methods to estimate
49 species trees and historical demographic parameters such as effective population sizes,
50 population divergence times, and gene flow holds that each DNA sequence locus is
51 “independent” from other sampled loci. This assumption is important because genealogical
52 histories (i.e., gene trees) of sampled loci are considered as true replicate samples depicting the
53 ancestry of a genome in these statistical analyses (Edwards & Beerli 2000; Arbogast et al. 2002;
54 Wakeley 2009). Indeed, the property of genealogical independence of loci confers benefits to
55 phylogenomic studies because larger numbers of independent loci enhance the accuracy and
56 precision of parameter estimates (Pluzhnikov & Donnelly 1996; Edwards & Beerli 2000;
57 Arbogast et al. 2002; Jennings & Edwards 2005; Felsenstein 2006; Lee & Edwards 2008; Smith
58 et al. 2013; Costa et al. 2016). Although the independent loci assumption is often mentioned in
59 coalescent-based studies, there is significant variation in how this assumption has been phrased
60 and interpreted.

61 We will now examine some examples taken from the literature, which show how
62 researchers have treated the independent loci assumption in phylogenomics (italics and bold are
63 mine). Arbogast et al. (2002) wrote: “*Indeed, the variance associated with estimates of*
64 *divergence time between recently diverged species can be minimized not by sequencing a large*
65 *number of sites per locus but by sequencing a large number of **independently segregating loci**;*”
66 Hudson & Coyne (2002): “*For results concerning multiple loci, we assume **statistical***
67 ***independence of the gene trees at different loci**;*” Yang (2002): “*It is assumed there is no*
68 *recombination within a locus and **free recombination between loci**;*” Hey & Nielsen (2004): “*A*
69 *key assumption of the method is that the **locus being studied has been evolving neutrally and***

70 *that it has been drawn at random from all loci, with respect to genealogical history;*” Bryant et
71 al. (2012): “*The genealogies for separate markers are conditionally independent given the*
72 *species tree;*” McCormack et al. (2012): “*Although it is increasingly feasible to sequence entire*
73 *genomes, identifying portions of the genome that are orthologous and **independently sorting** is*
74 *highly desirable from the perspective of analyses that take coalescent stochasticity into account;*”
75 Reilly et al. (2012): “*Our demographic parameter estimates may depend on the assumptions of*
76 *the IM model, which include loci **independently assort in meiosis;***” and lastly, O’Neill et al.
77 (2013) stated “*To maximize coverage of the genome and independence of loci, we chose loci that*
78 *ranged from approximately 200-650 bp in length, **were widely distributed across all 14 linkage***
79 *groups and were on average about 50 cM from other included loci on the Ambystoma linkage*
80 *map.*” As this brief survey shows, researchers have identified independent loci in at least two
81 different ways. In the first, independent loci are those that have independent genealogical
82 histories, whereas in the second independent loci are those that undergo Mendelian independent
83 assortment in meiosis. Several of the above bold-emphasized excerpts including “independently
84 segregating loci,” “free recombination between loci,” “independently sorting,” and loci being
85 “50 cM from other included loci,” presumably also refer to loci that undergo independent
86 assortment. A pair of intra-chromosomal loci that are separated by a map distance of at least 50
87 centimorgans (cM) are generally considered to be independently assorting in meiosis with
88 respect to each other. Thus, the independent loci assumption—as used in phylogenomic
89 studies—has evidently been conceptualized in at least two different ways. Studies that refer to
90 loci with independent genealogies are correctly encapsulating the independent loci assumption in
91 phylogenomics, whereas other studies are apparently confusing this assumption with the
92 independence assumption used in classical Mendelian genetics. However, it is unclear whether

93 the alternative interpretation (i.e., “independent assortment”) can also satisfy the independence
94 assumption in phylogenomics. Clarification of this inconsistency is important otherwise the
95 potential exists for some researchers to use incorrect or inefficient criteria for identifying
96 independent loci.

97 In order to precisely differentiate these two interpretations of the independence
98 assumption, we can think of each as a specific criterion: the first (hereafter criterion 1), considers
99 loci to be independent of other sampled loci if their genealogical histories are effectively
100 independent of each other, whereas under the second (hereafter criterion 2), sampled loci are
101 independent of each other if they undergo independent assortment. Criteria 1 and 2 are
102 equivalent when considering two loci found on different chromosomes—just as loci found on
103 different chromosomes will undergo independent assortment, such loci will also have
104 independent gene trees (Wakeley 2009). However, these criteria differ from each other regarding
105 the identification of genealogically independent loci found on the *same* chromosomes. While
106 criterion 1 takes into account both the long-term effects of homologous recombination and
107 effective population size (N_e), criterion 2 only considers the effects of homologous
108 recombination (i.e., no demographic component). Thus, regarding loci found on the same
109 chromosomes, these criteria are fundamentally different from each other and this difference has
110 important implications for phylogenomic studies.

111 Advances in next generation sequencing are enabling researchers to obtain phylogenomic
112 datasets consisting of hundreds to thousands of targeted loci via in-solution sequence capture
113 methods (e.g., Gnirke et al. 2009; Faircloth et al. 2012; Lemmon et al. 2012; Meikeljohn et al.
114 2016) or whole-genome sequencing (e.g., Jarvis et al. 2014). Thus, a need exists for practical
115 methods that can identify loci that likely meet the independence assumption otherwise large

116 genome-wide datasets may inadvertently include pseudoreplicated loci (Costa et al. 2016). One
117 approach that has been used to identify putatively independent loci in samples has been to use
118 complete genome data in conjunction with an *a priori* “distance threshold,” which represents the
119 minimum intra-chromosomal “distance” between two sampled loci that are presumed to have
120 independent gene trees. These distances have been in the form of physical distances in units of
121 base pairs or “bp” (e.g., Sachidanandam et al. 2001; Leaché et al. 2015; Costa et al. 2016) or a
122 recombination distance in units of cM (e.g., O’Neill et al. 2013). In other studies, researchers
123 evaluated their datasets in an *a posteriori* manner by observing that sampled loci were separated
124 from each other by vast intra-chromosomal distances (e.g., > 1 Mb) and therefore likely satisfied
125 the independence assumption (e.g., McCormack et al. 2012). However, only the studies of Costa
126 et al. (2016) and O’Neill et al. (2013) used threshold distances based on stated objective criteria:
127 the former study implicitly invoked criterion 1, whereas the latter invoked criterion 2.
128 Nonetheless, all studies that have made some effort to ensure that their multi-locus datasets were
129 largely compliant with the independent loci assumption have helped move the field of
130 phylogenomics forward.

131 Here, I evaluate these criteria for delimiting independent loci using empirical examples.
132 As we will see, if sufficient data are available, then both criteria can be used to identify
133 independent loci in a sample. However, we will also see that one of these two criteria is likely to
134 be far too conservative for use in many phylogenomic studies.

135

136 **Materials and Methods**

137 To illustrate the relative utility of each criterion for delimiting independent loci in eukaryotic
138 genomes, both criteria are examined using genetic and genomic information available for the

139 common fruit fly (*Drosophila melanogaster*) and North American Tiger Salamanders
140 (*Ambystoma tigrinum*). Hudson & Coyne (2002) developed a theoretical framework that can be
141 used to identify independent loci under criterion 1. These authors referred to independent loci
142 whose gene trees are statistically independent of each other as being *independent genealogical*
143 *units* or “IGUs,” which they defined as “*the number of genomic segments whose passage to*
144 *monophyly is nearly independent of that for all other segments*” (Hudson & Coyne 2002).
145 Furthermore, these authors derived a formula for estimating the total number of IGUs in a
146 genome, which is shown here in the following general form found in Costa et al. (2016):

$$147 \qquad \qquad \qquad \text{IGUs} = 4N_e c / 1,000 \qquad \qquad \qquad (1)$$

148 whereby N_e is effective population size and the c is the per generation recombination rate. As
149 mentioned earlier, criterion 1 contains a demographic component and this aspect is plainly
150 evident in formula (1), which shows that N_e plays a role in determining the number of loci with
151 effectively independent genealogies. Thus, for a given recombination rate, large N_e values
152 translate to more IGUs per genome than smaller N_e values and vice-versa. Hudson & Coyne
153 (2002) estimated the number of IGUs in the *D. melanogaster* genome, which is based on a
154 genetic map length of ~287 cM and N_e of 10^6 for this species (see Results and Discussion).

155 The number of IGUs in the *A. tigrinum* genome under criterion 1 was estimated using the
156 genetic linkage map for the Mexican Axolotl (*A. mexicanum*), which is 5,251 cM in length
157 (Smith et al., 2005). However, in order to use formula (1), an estimate of N_e must also be
158 supplied, which is problematic because North American Tiger Salamanders have widely varying
159 N_e depending on the species. For example, Wang et al. (2011) found that California Tiger
160 Salamanders (*A. californiense*) had exceedingly low N_e of 11-64, which may be explained by
161 population bottlenecks or pond sizes. In contrast, Church et al. (2003), who used mitochondrial

162 DNA, estimated the effective number of females (N_f) in Eastern Tiger Salamanders (*A. tigrinum*)
163 to be 134,000-144,000. Because autosomal loci have 4-fold higher N_e than mitochondrial loci
164 (Wilson et al. 1985), N_e for autosomal loci in these salamanders are likely higher. Owing to this
165 wide-ranging variation in N_e across North American *Amybystoma* species and populations it is
166 difficult to know which N_e value should be inserted into formula (1) above. However, as these
167 salamanders currently have a continental-wide distribution, they may have had more genetic
168 connectivity among populations in the past. Therefore, N_e values of 10^3 - 10^5 appear reasonable
169 for our present purpose, particularly in light of the recent phylogenomic study of this entire
170 species complex by O'Neill et al. (2013).

171 Criterion 2 (independent assortment) only requires a genetic linkage map for the study
172 species or group and thus it is simpler to use than criterion 1. Thus, given the map length of ~287
173 cM for the *D. melanogaster* genome (Hudson & Coyne 2002), it was straightforward to estimate
174 the number of independent loci under under criterion 2. O'Neill et al. (2013) were evidently the
175 first researchers to use the independent assortment criterion to select their phylogenomic loci.
176 Using the *A. mexicanum* linkage map these authors developed 95 PCR-based loci taken from all
177 14 linkage groups and ensured that no two loci were closer than 50 cM apart on the same
178 chromosomes (O'Neill et al. 2013). In the current study, the total number of independent loci in
179 the *Ambystoma* genome under criterion 2 was estimated.

180

181 **Results and Discussion**

182 Under criterion 1, the fruit fly genome contains approximately 11,500 IGUs (Hudson & Coyne
183 2002). Thus, given a genome size of ~143 Mb for this species (NCBI 2016), we would expect,
184 on average, to encounter one IGU or independent locus every ~12,500 bp along its

185 chromosomes. However, this type of distance threshold should be regarded as a rough estimate
186 because local recombination rates and N_e vary across genomes (Costa et al. 2016). Nonetheless,
187 this threshold value still provides us with some means for deciding whether any given nearest-
188 neighbor pair of loci found on the same chromosome may be genealogically independent of each
189 other or not. What are the comparable estimates under criterion 2? If we assume that loci
190 separated by 50 cM on the same chromosomes are independent from each other, then we would
191 conclude that there are only six independent loci in this genome. In reality, however, there must
192 be at least seven IGUs because there must be one IGU for each of the seven chromosomes in the
193 *D. melanogaster* genome. This means we would expect to see one independent locus per 20 Mb
194 in the genome. In summary, the number of independent loci under criteria 1 and 2 are ~11,500
195 and seven, respectively, while the inter-locus distance thresholds are ~12.5 kb and 20 Mb,
196 respectively. Clearly, criterion 2 is far too conservative to be of practical use for fruit flies.

197 Using equation (1), the number of IGUs in the tiger salamander genome is equal to
198 $[(4)(1,000)(5,251 \text{ cM})(0.01 \text{ cross-overs per generation})]/1,000 = 210$ IGUs. If the long-term N_e is
199 instead assumed to be larger at 10^5 , then the number of IGUs increases a hundred-fold to 21,000.
200 With these IGU estimates and knowing that the genome of *A. mexicanum* is 354 Mb in size
201 (NCBI 2016), we can expect to see one IGU every 17 kb to 1.7 Mb depending on whether the
202 assumed N_e value is 10^5 or 10^3 , respectively. Under criterion 2, there are 105 IGUs in the
203 *Ambystoma* genome, which translates to about one independent locus per 3.4 Mb, on average.
204 Although the estimated number of independent loci in the tiger salamander genome under
205 criterion 2 is by no means a small number of loci for a phylogenomic dataset, it is still
206 substantially smaller than the number of loci that would be obtained using criterion 1 even if a
207 low N_e were to be assumed.

208 The fruit fly and tiger salamander examples demonstrate that the independent assortment-
209 based criterion for identifying genealogically independent loci is overly stringent and would
210 therefore unfairly restrict researchers to using fewer independent loci than would be permitted
211 under the genealogical-based criterion. Accordingly, for evolutionary studies involving multi-
212 locus coalescent analyses it is recommended that researchers use, whenever possible, the
213 criterion of genealogical independence for independent loci (or SNPs). Although criterion 1 is
214 more difficult to implement than criterion 2 owing to its requirement of an estimate of N_e , it
215 offers a promising approach for elucidating appropriate physical distance thresholds between
216 independent loci in genomes. This, in turn, should allow researchers to generate phylogenomic
217 datasets with the maximum number of genealogically independent loci or SNPs.

218

219 **References**

- 220 Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB. 2002. Estimating divergence
221 times from molecular data on phylogenetic and population genetic timescales. *Annual*
222 *Review of Ecology and Systematics*, 1:707-40.
- 223 Bryant D, Bouckaert R, Felsenstein J, Rosenberg NA, RoyChoudhury A. 2012. Inferring species
224 trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent
225 analysis. *Molecular Biology and Evolution*, 29:1917-1932.
- 226 Church SA, Kraus JM, Mitchell JC, Church DR, Taylor DR. 2003. Evidence for multiple
227 Pleistocene refugia in the postglacial expansion of the eastern tiger salamander,
228 *Ambystoma tigrinum tigrinum*. *Evolution*, 57:372-383.
- 229 Costa IR, Prosdocimi F, Jennings WB. 2016. In silico phylogenomics using complete genomes: a
230 case study on the evolution of hominoids. *Genome Research* doi: 10.1101/gr.203950.115.

- 231 Edwards SV, Beerli P. 2000. Perspective: gene divergence, population divergence, and the
232 variance in coalescence time in phylogeographic studies. *Evolution*, 54:1839-54.
- 233 Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. 2012.
234 Ultraconserved elements anchor thousands of genetic markers spanning multiple
235 evolutionary timescales. *Systematic Biology*, 61:717-726.
- 236 Felsenstein J. 2006. Accuracy of coalescent likelihood estimates: do we need more sites, more
237 sequences, or more loci?. *Molecular Biology and Evolution*, 23:691-700.
- 238 Gnirke A, Melnikov A, Maguire J, Rogov P, LeProust EM, Brockman W, Fennell T,
239 Giannoukos G, Fisher S, Russ C, Gabriel S, Jaffe DB, Lander ES, Nusbaum C. 2009.
240 Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted
241 sequencing. *Nature Biotechnology*, 27:182-189.
- 242 Hey J, Nielsen R. 2004. Multilocus methods for estimating population sizes, migration rates and
243 divergence time, with applications to the divergence of *Drosophila pseudoobscura* and
244 *D. persimilis*. *Genetics*, 167:747-760.
- 245 Hudson RR, Coyne JA. 2002. Mathematical consequences of the genealogical species concept.
246 *Evolution*, 56:1557-1565.
- 247 Jarvis ED, Mirarab S, Aberer AJ, Li B, Houde P, Li C, Ho SY, Faircloth BC, Nabholz B,
248 Howard JT, Suh A. 2014. Whole-genome analyses resolve early branches in the tree of
249 life of modern birds. *Science*, 346:1320-31.
- 250 Jennings WB, Edwards SV. 2005. Speciation history of Australian Grass Finches *Poephila*
251 inferred from thirty gene trees. *Evolution*, 59:2033-2047.
- 252 Leaché AD, Chavez AS, Jones LN, Grummer JA, Gottscho AD Linkem CW. 2015.
253 Phylogenomics of Phrynosomatid lizards: conflicting signals from sequence capture

- 254 versus restriction site associated DNA sequencing. *Genome Biology and Evolution*,
255 7:706-719.
- 256 Lee JY, Edwards SV. 2008. Divergence across Australia's Carpentarian barrier: statistical
257 phylogeography of the Red-backed Fairy Wren (*Malurus melanocephalus*). *Evolution*,
258 62:3117-3134.
- 259 Lemmon AR, Emme SA, Lemmon EM. 2012. Anchored hybrid enrichment for massively high-
260 throughput phylogenomics. *Systematic Biology*, p.sys049.
- 261 McCormack JE, Faircloth BC, Crawford NG, Gowaty PA, Brumfield RT, Glenn TC. 2012.
262 Ultraconserved elements are novel phylogenomic markers that resolve placental mammal
263 phylogeny when combined with species-tree analysis. *Genome Research*, 22:746-754.
- 264 Meiklejohn KA, Faircloth BC, Glenn TC, Kimball RT, Braun EL. 2016. Analysis of a rapid
265 evolutionary radiation using ultraconserved elements: evidence for a bias in some
266 multispecies coalescent methods. *Systematic Biology*, p.syw014.
- 267 NCBI (National Center for Biotechnology Information) Genome Database. Retrieved 20 July
268 2016.
- 269 O'Neill EM, Schwartz R, Bullock CT, Williams JS, Shaffer HB, Aguilar-Miguel X, Parra-Olea
270 G, Weisrock DW. 2013. Parallel tagged amplicon sequencing reveals major lineages and
271 phylogenetic structure in the North American tiger salamander (*Ambystoma tigrinum*)
272 species complex. *Molecular Ecology*, 22:111-129.
- 273 Pluzhnikov A, Donnelly P. 1996. Optimal sequencing strategies for surveying molecular genetic
274 diversity. *Genetics*, 144:1247-1262.
- 275 Reilly SB, Marks SB, Jennings WB. 2012. Defining evolutionary boundaries across parapatric

- 276 ecomorphs of Black Salamanders (*Aneides flavipunctatus*) with conservation
277 implications. *Molecular Ecology*, 21:5745-5761.
- 278 Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin
279 JC, Mortimore BJ, Willey DL, et al. 2001. A map of human genome sequence variation
280 containing 1.42 million single nucleotide polymorphisms. *Nature*, 409:928-933
- 281 Smith BT, Harvey MG, Faircloth BC, Glenn TC, Brumfield RT. 2013. Target capture and
282 massively parallel sequencing of ultraconserved elements for comparative studies at
283 shallow evolutionary time scales. *Systematic Biology*, DOI:10.1093/sysbio/syt061.
- 284 Smith JJ, Kump DK, Walker JA, Parichy DM, Voss SR. 2005. A comprehensive expressed
285 sequence tag linkage map for tiger salamander and Mexican axolotl: enabling gene
286 mapping and comparative genomics in *Ambystoma*. *Genetics*, 171:1161-1171.
- 287 Wakeley J. 2009. *Coalescent theory: an introduction* (Vol. 1). Greenwood Village: Roberts &
288 Company Publishers.
- 289 Wang IJ, Johnson JR, Johnson BB, Shaffer HB. 2011. Effective population size is strongly
290 correlated with breeding pond size in the endangered California tiger salamander,
291 *Ambystoma californiense*. *Conservation Genetics*, 12:911-920.
- 292 Wilson AC, Cann RL, Carr SM, George M, Gyllensten UB, Helm-Bychowski KM, Higuchi RG,
293 Palumbi SR, Prager EM, Sage RD, Stoneking M. 1985. Mitochondrial DNA and two
294 perspectives on evolutionary genetics. *Biological Journal of the Linnean Society*, 26:375-
295 400.
- 296 Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in hominoids using
297 data from multiple loci. *Genetics*, 162:1811-1823.