A 4-lineage statistical suite to evaluate the support of large-scale retrotransposon insertion data to reconstruct evolutionary trees

Churakov G*[§1], Kuritzin A[§2], Chukharev K[3], Zhang F[1], Wünnemann F[4], Ulyantsev V[3], and Schmitz J*[1]

[1]Institute of Experimental Pathology (ZMBE), University of Münster, Münster, Germany

[2]Department of System Analysis, Saint Petersburg State Institute of Technology, St. Petersburg, Russia

[3]Information Technologies, Mechanics and Optics, University Saint Petersburg, St. Petersburg, Russia

[4]Montreal Heart Institute, Universite de Montreal, Quebec, Canada

***Corresponding authors:** churakov@uni-muenster.de, jueschm@uni-muenster.de

[§]These authors contributed equally to this work.

Running title: Phylogenomic 4-lineage retrotransposon statistics

Keywords: ancestral incomplete lineage sorting, ancestral hybridization, polytomy, retrophylogenomics, KKSC, 4-lineage (4-LIN) insertion polymorphism, asymmetric and symmetric distribution

Abstract

Retrophylogenomics makes use of genome-wide retrotransposon presence/absence insertion patterns to resolve questions in phylogeny and population genetics. In the genomics era, evaluating high-throughput data requires the associated development of appropriately powerful statistical tools. The currently used KKSC 3-lineage statistical test for evaluating the significance of data is limited by the number of possible tree topologies it can assess in one step. To improve on this, we have now extended the analysis to simultaneously compare 4-lineages, which now enables us to evaluate ten distinct presence/absence insertion patterns for 26 possible tree topologies plus 129 trees with different incidences of hybridization. Moreover, the new tool includes statistics for multiple ancestral hybridizations, ancestral incomplete lineage sorting, bifurcation, and polytomy. The test is embedded in a user-friendly web R-application (http://retrogenomics.uni-muenster.de:3838/hammlet/) and is available for use by the general scientific community.

Introduction

Known to be virtually homoplasy-free, retrotransposon phylogenetic presence/absence markers require careful manual inspections to be certain of the true orthology of their loci and evaluations of different potential tree topologies. Statistical tests evaluate whether the number of shared insertions inherited from a common ancestor compared with the absence of such elements in other species is significant to support their relatedness. However, occasionally, such presence/absence markers appear to support contradictory phylogenetic tree topologies. Such potential conflicts may be evoked by (1) extremely rare cases of

parallel retroposon insertions in unrelated species (about 0.01% of diagnostic markers), (2) even rarer exact deletions in one or more related species (about 0.001% of diagnostic markers) (Doronina et al., 2019), or (3) most influentially, ancestral hybridization/introgression and incomplete lineage sorting (ILS), the extents of which can differ for diverse taxonomic groups (Doronina et al., 2019). ILS is sometimes associated with the evolution of species on short internodal phylogenetic branches. These short branches are indicative of short periods between speciation, too short in fact for all inserted retrotransposons to have been fixed in all related species of a particular group before the next speciation occurred. Thus, some members of the group have the insertion and other not, even though they still belong to the same phylogenetic group (Kuritzin et al., 2016).

Waddell et al. (2001) developed the first statistical test to evaluate the probability of a particular presence/absence marker insertion pattern supporting a prior hypothesis of relatedness against polytomy. However, the Waddell test returns p-values for only up to 5 marker combinations and is therefore no longer suitable for present day genome analysis with hundreds or even thousands of markers. Kuritzin et al. (2016) developed the 3-lineage KKSC statistics, which introduced multidirectional analyses that can evaluate the presence/absence patterns of phylogenetic markers present in all potential tree topologies of three species without a prior hypothesis. It also includes a less powerful, one-sided test, in which the phylogenetic markers found in one reference lineage are screened for their presence/absence in other species (http://retrogenomics.uni-muenster.de:3838/KKSC_significance_test/). KKSC also includes a simple consideration of ancestral hybridization based on the symmetric or asymmetric distribution of phylogenetic markers. The KKSC statistics, based on a probability calculation, are limited however to

evaluating the evidence for the interrelatedness of only 3 lineages and their three corresponding potential tree topologies. Furthermore, they insist on a combination of sets of 3 taxa to evaluate larger trees. By changing to a maximum likelihood calculation and expanding the evaluation to handle 4 lineages, thereby testing 155 different tree topologies, 129 of which with varying extents of multiple hybridization scenarios, most phylogenetic questions can be accessed directly or in a combination of 4 species. Use of the 4-lineage (4-LIN) test requires multidirectional screening for phylogenetic markers starting from four different reference species and considers 10 distinct informative presence/absence patterns for each phylogenetic marker.

Methods

Following our previous development (Kuritzin et al., 2016), we extended the diffusion approximation of Kimura (1955ab) to four lineages (Doronina et al., 2017a). Based on presence/absence patterns of inserted retrotransposon, we developed new statistical criteria using log-likelihood ratio tests to identify the most likely supported phylogenetic tree from 155 four-lineage topologies that include multiple hybridization scenarios. To apply the 4-LIN statistical test, retrotransposon presence/absence patterns must first be derived from species representing four lineages. To enable an unbiased screening, we recommend to start with qualitatively equal genome assemblies (covering equal percentages of the genomes). There is no limit to the number of phylogenetic markers that can be evaluated. The input data for statistical comparison of the presence/absence patterns across four lineages are 10 diagnostic marker combinations (order of diagnostic patterns: -+++, --++, -+-+, -++-, +-++, +--+, +-+-, ++-+, ++--, +++-).

Model assumption

If we consider four lineages A_1, A_2, A_3, A_4 sharing common ancestry, we can describe each branch B_k as an isolated population of individuals over an evolutionary time interval $t \in \Delta_k$ with an effective population size $N_k(t)$. We use the fusion model to record hybridization, whereby two separated ancestral populations reproduce to form a new branch (Kuritzin et al., 2016). It should be noted that both hybridization and introgression can be described in terms of the fusion model but cannot be kept apart by experimental data. We can consider two possible basic models 2H1 and 2H2 of speciation for four lineages, including each of two potential hybridization events (Fig. 1).

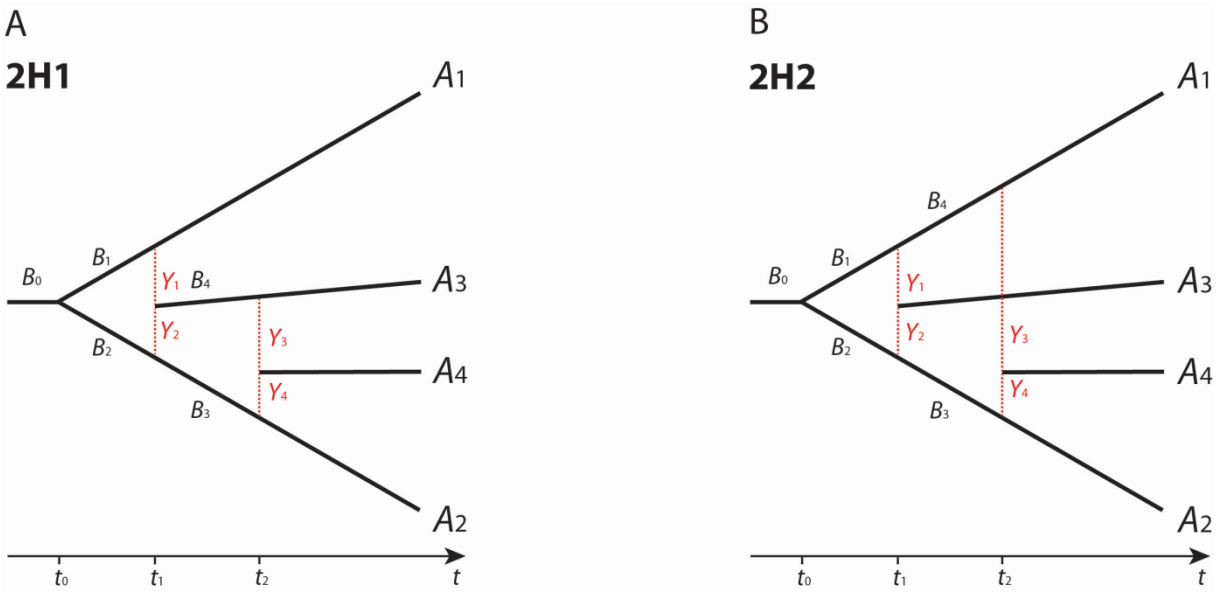

Figure 1. Two basic evolutionary models involving the four lineages A_1, A_2, A_3, A_4. Black lines indicate different lineages and red vertical dotted lines indicate hybridization events. A) Basic model 2H1: sequential hybridization, B) Basic model 2H2: parallel hybridization. The proportions of two fused subpopulations forming a new population at time $t = t_1$ are denoted y_1 and y_2 ($y_1 + y_2 = 1$). At the second fusion point ($t = t_2$), the sub-population proportions are denoted y_3 and y_4 ($y_3 + y_4 = 1$).

For each of these two hybridization scenarios, we can derive 24 permutations of the four lineages A_1, A_2, A_3, A_4. However, due to symmetry, the number of rearrangements for 2H2 can be reduced to 12 permutations (see Fig. 2).

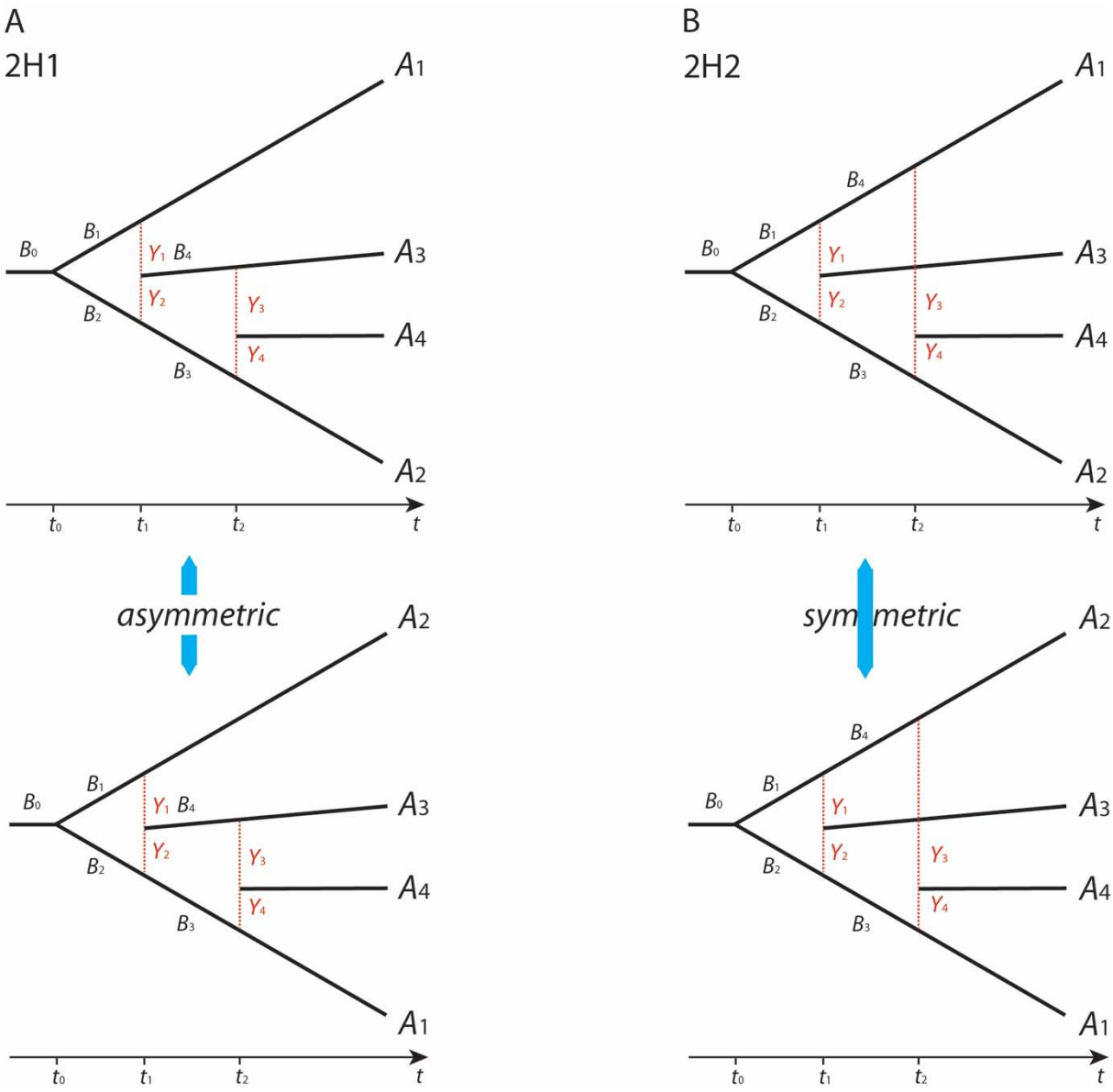

Figure 2. Asymmetric and symmetric permutations. A) Asymmetric permutation of model 2H1. Permutations of $A_{1,2,3,4}$ and $A_{2,1,3,4}$ lead to different trees. A_3 is the result of hybridization between A_1 and A_2 for both the upper and lower permutations, respectively. However, A_4 is the result of the hybridization of A_2 and A_3 in the upper tree and of A_1 and A_3 in the lower tree. B) The symmetric permutation of model 2H2 leads to identical phylogenetic trees ($A_{1,2,3,4}$ and $A_{2,1,3,4}$), taking in account the exchange of the value of γ_1 by γ_2 and γ_3 by γ_4.

For a neutral insertion of retroelements, in generation *t* of the branch B_k we consider ten different events $\omega_{i,j}$ ($1 \le i \le j \le 4$) for four lineages where:

in $\omega_{i,i}$ a retroelement is absent in the orthologous locus of lineage A_i but present in the other three lineages;

in $\omega_{i,j}$ a retroelement is absent in the orthologous loci of lineages A_i and A_j ($i \neq j$) but present in the other two lineages.

Assuming that the probability of new insertions for each individual of a population is small and the effective population size is large, we can conclude that the total number of retroelement insertions with the property $\omega_{i,j}$ are independent, Poisson-distributed random variables with parameters $a_{i,j}$.

We analyzed both of these models to find $a_{i,j}$ (see Appendix 1, S1.3.1.-S1.3.30). The calculations were carried out similarly to those in Kuritzin et al. (2016) under the assumption that the effective population sizes $N_k(t)$ at the corresponding intervals are constant. Hence, denoting $r_k = \frac{N_k}{N_0}$ $(1 \le k \le 4)$ and fixing r_k, we can write $a_{i,j} = n_0 \bar{a}_{i,j}$, with $\bar{a}_{i,j}$ depending on the four values $T_1, T_3, \gamma_1, \text{and } \gamma_3$ (noting that $\gamma_2 = 1 - \gamma_1$ and $\gamma_4 = 1 - \gamma_3$), where $T_1 = \frac{t_1 - t_0}{2N_0}, and\ T_3 = \frac{t_2 - t_1}{2N_0}$, reflecting the periods from the first split of two branches from an ancestral population to the first hybridization event, and from the first hybridization event to the second hybridization event, and where n_0 is a stray parameter (see Appendix 1, S1.4.1-S1.4.10).

Model investigations

For modeling different tree variants, we need to fix the model parameters T_1, T_3, γ_1, γ_3 to their boundaries. It should be mentioned that fixing T_1 and T_3 to the boundary may lead to

branch lengths of zero (T_1=0 and T_3=0), and that γ_1 and γ_3 can be fixed to the two extremes γ=0 or γ=1, leading to different models. When a branch is of zero length (T=0), in most cases γ is undefined (changing γ does not change the model).

For further discussion we introduce the "TTgg" nomenclature of models. E.g. the 2H1 model we annotate as H1:TTgg, where each T or g reflects the status of parameter: first "T" reflects T_1, second "T" corresponds to T_3. The first "g" shows the status of γ_1, and the second "g" represents γ_3; the prefix (H1 or H2) points to basic model (2H1 or 2H2).

To switch between different models, we can vary specific parameters, e.g., adjusting γ_3 to 0 will switch H1:TTgg (2H1 model with double hybridization; see Appendix 1, Table 1) to H1:TTg0 (1H1 model). Or, adjusting γ_3 to 1 will switch the same model to H1:TTg1 (1H4 model).

From the model 2H1, which has 24 different permutations of its four lineages, we can create 20 different models (each with 24 permutations), and from the model 2H2, which has 12 permutations, we can derive 22 different models, each with 12 permutations (see Appendix 1, Table 1). Thus, we can derive a total of 744 variants ($24 \times 20 + 12 \times 22$) of trees (models with fixed orders of species; e.g., 2H1:1234 or H1:TTgg:1234). Excluding symmetric options (see Fig. 2) and duplicated models with identical biological meaning reduce this to a set of 15 models and 155 unique trees (see Appendix 1, Table 2 for a non-redundant list of model variants and permutations). Assuming models 2H1 and 2H2 (Fig. 1), we can sort diagnostic cases into five groups depending on the number of fixed parameters (Figures 3, 4, 5, 6).

The first group, designated θ0, includes one tree with all four parameters fixed, reflecting polytomy P (0000, one permutation, Fig. 3A). The second, Group θ1 comprises two models with one released parameter. The first model in group θ1 is PT, where the first

diversification point forms a polytomy and the second diversification point yields two new branches (H1:0Tn1, six permutations, Fig. 3B). Model TP is derived from P by releasing the first period T_1. In this model, a polytomy appears after a second diversification point (H1:T001, four permutations, Fig. 3C).

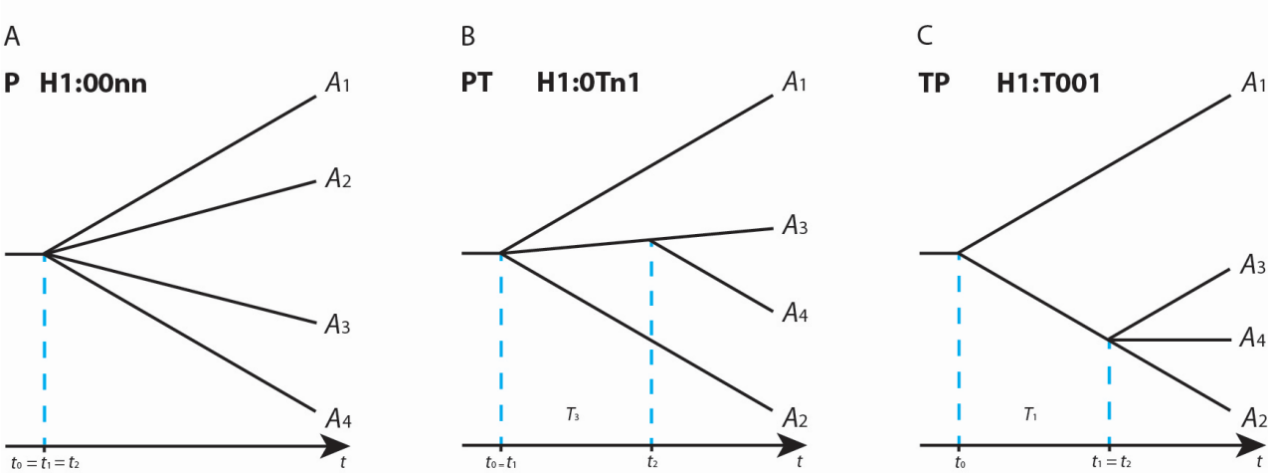

Figure 3. Groups θ0 and θ1. The arrows below the trees indicate the times (t) of splits, with $t_{0,1,2}$ being specific events. t_0 indicates the initial split, t_1 the first hybridization or split, and t_2 the second split or hybridization event. t_0=t_1, etc., signifies that two events appeared simultaneously. The names of models (and aliases) are shown above the graph. A1,2,3,4 indicate the order of lineages in the first permutation (1234). T_3 and T_1 indicate different time periods and are shown only if they are not fixed in the respective model. A) Polytomy. B) The first diversification (at t_0=t_1) is polytomy followed by a dichotomic split. C) The first diversification (at t_0) is dichotomy followed by polytomy (at t_1=t_2).

Group θ2 comprises a hybridization-polytomy tree, two binary tree models, wherein two parameters are released, and lastly two binary tree models. Model 1HP1 represents hybridization-polytomy derived from the model PT with a second hybridization coefficient γ_3 (H1:0Tng, 12 permutations, Fig. 4A). Model 1HP2 is derived from TP by releasing the first hybridization coefficient γ_1, which indicates that hybridization and branch divergence happen simultaneously (H1:T0g1, 12 permutations, Fig. 4B, see also Doronina et al. (2017a). Model 1HP3 shows hybridization combined with diversification of hybridization branches

(H1:T0g1, six permutations, Fig. 4C). The first binary model T1 represents two independent diversifications following the first split (H1:TT10, six permutations, Fig. 4D), while the second model T2 represents three consequent diversifications of an ancestral branch (H1:TT01, 12 permutations, Fig. 4E).

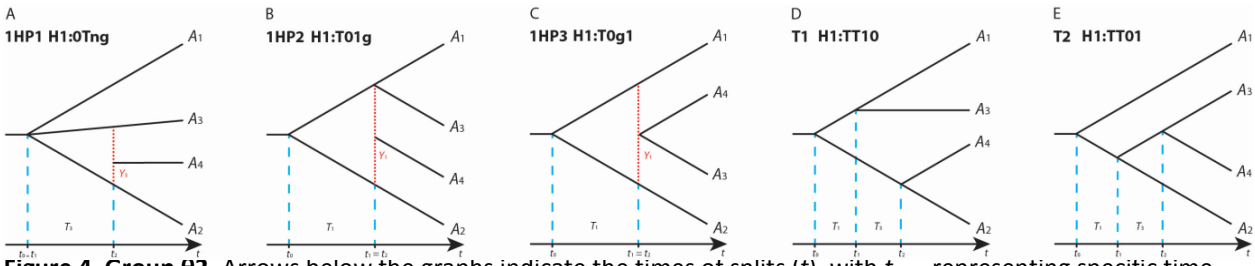

Figure 4. Group θ2. Arrows below the graphs indicate the times of splits (t), with $t_{0,1,2}$ representing specific time events. t_0 indicates the initial splitting point, t_1 the first hybridization or split event, and t_2 indicates a second split/hybridization event. t_0=t_1, etc., indicate simultaneous events. Models and aliases are labeled above the tree graphics. $A_{1,2,3,4}$ indicate the order of lineages in the first permutation (1234). T_1 and T_3 indicate different time periods and γ_1 and γ_3 the hybridization coefficients. The last two sets of parameters are shown only if they are not fixed. A) three branches (A_1, A_3, A_2) were the result of an initial diversification and an incidence of hybridization (at t_2) between two of these branches. B) hybridization took place simultaneously with the second diversification (at $t_1 = t_2$). C) The resulting hybridization (at $t_1 = t_2$) immediately diversified into two branches (A_4, A_3). D) Tree variant with two independent, final diversifications (at t_1, t_2). E) Option with two subsequent last diversifications.

Group θ3 comprises five specific models, in which three parameters are freed and one is fixed. One with two hybridizations at the same time after the first split of 2HP (H2:T0gg, six permutations, see Fig. 5A), and four models with single hybridizations. Model 1H1 represents a situation where hybridization follows the initial split before a second one (H1:TTg0, 12 permutations, see Fig. 5B). Model 1H2 represents hybridization after the initial and first splits and between distant branches (H1:TT1g, 24 permutations, see Fig. 5C). Model 1H3 contains a hybridization between sister branches following the initial and first splits

(H1:TT0g, 12 permutations, Fig. 5D). Model 1H4 shows hybridization after the initial split and a subsequent second split after hybridization (H1:TTg1, six permutations, Fig. 5E).

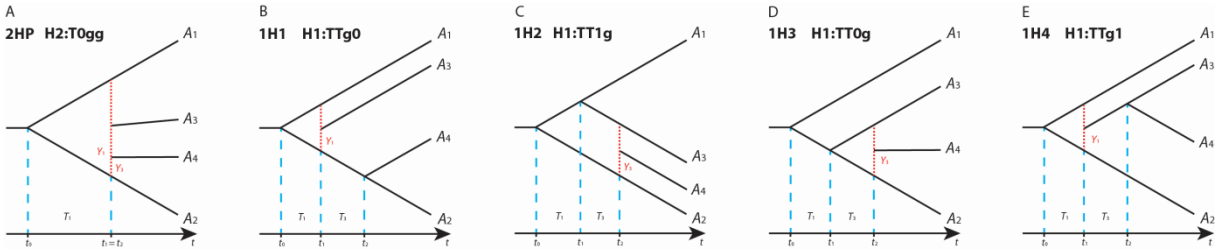

Figure 5. Group θ3. Arrows below the trees indicate the times of splits (t), with $t_{0,1,2}$. t_0 denotating the initial splitting point, t_1 the first hybridization or split event, and t_2 the second hybridization or split event. t_1=t_2 indicates simultaneous events. Names (and aliases) of models are presented above the trees. $A_{1\text{-}4}$ indicate the order of lineages in the first permutation (1234), and T_1 and T_3 indicate different time periods, γ_1, and γ_3 are the hybridization coefficients. Parameters in the last two trees are only shown if they are not fixed. A) two simultaneous hybridizations, B) hybridization at t_1. C) hybridization after the first split (between the youngest and most distant branches). D) hybridization after the first split (between the two most budding sister branches), E) Branch derived from hybridization after initial split represents the last diversification.

Group θ4 includes 24 permutations of 2H1 (H1:TTgg, Fig. 6A, root model) and 12 permutations from 2H2 (H2:TTgg, Fig. 6b, root model), in which all parameters are free to change.

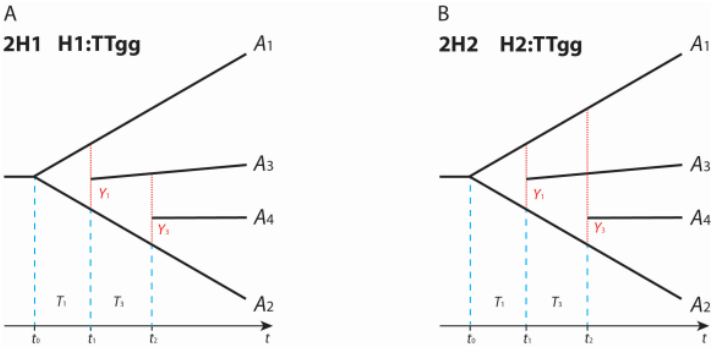

Figure 6. Group θ4. Arrows indicate the times (t) of splits with events t_0, t_1, t_2. Here, t_0 denotes the initial split, t_1 marks the first hybridization event, and t_2 the second. Model names (2H1, 2H2) and aliases (H1:TTgg, H2:TTgg) are above the trees. $A_{1\text{-}4}$ denote the order of lineages in the first permutation (1234), and T_1 and T_3 indicate different time periods, while γ_1, and γ_3 are the hybridization coefficients. A) two consecutive hybridizations, B) two parallel hybridizations.

Finally, for the five groups (θ0 - 1 tree, θ1 - 10 trees, θ2 - 48 trees, θ3 - 60 trees, θ4 - 36 trees), we calculated the optimized likelihood values. Using this statistical approach, we can then determine the most probable tree for each group that best fits the distribution of phylogenetic marker insertions in the species examined.

3. Statistical testing

The different parameters described above contribute specifically to the likelihood estimation, even if none of them have equal influence on all of the models. The model parameters are estimated using the maximum likelihood method. Based on the empirical data x, the likelihood function $l(x;\theta)$ is calculated for each of the 155 trees as a function with unknown parameters θ; these values are then used to estimate the parameters. In the next step, for each group (θ0, θ1, θ2, θ3, θ4) the tree with the highest likelihood function $L(x|\theta_j) = \max_{\theta\in\theta j} l(x;\theta)$ is selected. Because each of the trees of group θj is a particular case of some model of group $\theta_{(j+1)}(0 \leq j \leq 3)$, then $L(x|\,\theta\,4) \geq L(x|\,\theta\,3) \geq L(x|\,\theta\,2) \geq L(x|\,\theta\,1) \geq L(x|\,\theta\,0)$ (see Appendix 1, S1.5.1-S1.5.4).

We propose two statistical algorithms, reverse and stepwise, to find the true tree (Fig. 7).

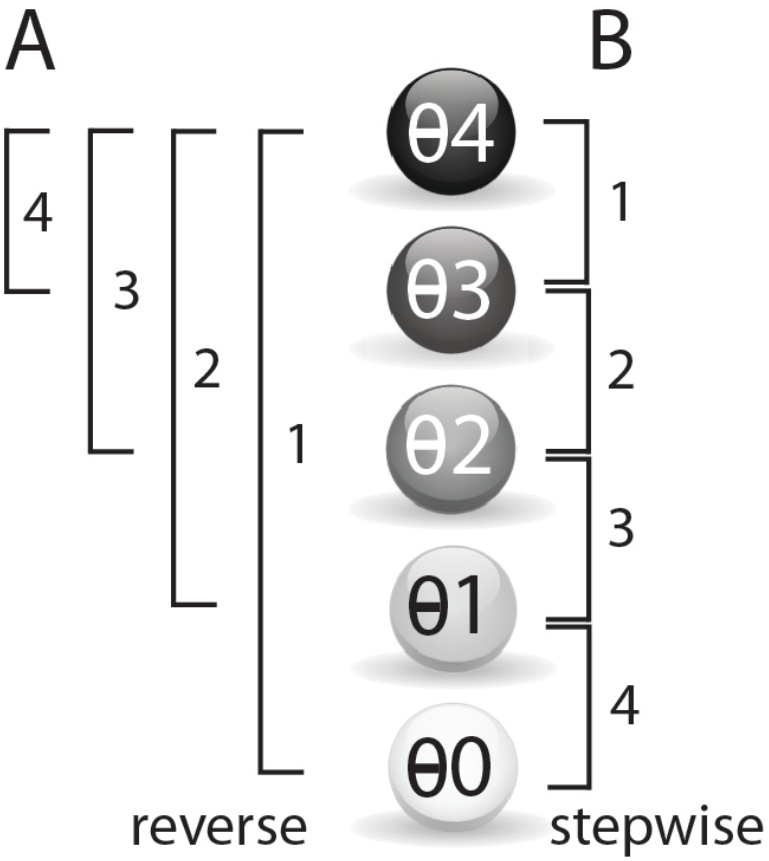

Figure 7. Statistical strategies. Colored circles represent the 5 groups of models from which the best tree estimation (corresponding to maximum likelihood) is derived. Brackets indicate the order of performed comparisons. The lengths of the brackets denote the degrees of freedom for chi-square. A) In the reverse method, the degrees of freedom (i.e., bracket lengths) for chi-square decrease from 4 to 1, B) In the stepwise method, the degrees of freedom (denoted by bracket length) for chi-square is stable and equivalent to one.

Reverse method:

As a measure of the likelihood of the tree Θ_j, we take the log-likelihood ratio $\lambda_j^4(x) = -2\log\frac{L(x|\Theta_j)}{L(x|\Theta_4)}$.

First, we calculate both the likelihood ratio $\lambda_0^4(x)$, and P-value $p_0^4(x) \approx \mathrm{P}\left(\chi_4^2 \geq \lambda_0^4(x)\right)$ - that can be extracted from the chi-square distribution. Here χ_k^2 denotes a random variable distributed according to chi-square with the number of degrees of freedom set to *k*.

If $p_0^4(x) > \alpha$ - with the predetermined significance level (e.g., $\alpha = 0.05$), we have no reason to reject Θ_0 (polytomy); if no polytomy, we proceed to test the best tree from Θ_1.

Thus, we calculate both the likelihood ratio $\lambda_1^4(x)$ and P-value $p_1^4(x) \approx \mathrm{P}\left(\chi_3^2 \geq \lambda_1^4(x)\right)$. If $p_1^4(x) > \alpha$, then the best tree of group Θ_1 is accepted; if not, we proceed to test the best tree from Θ_2.

The comparisons are repeated until $p_j^4(x) \approx \mathrm{P}\left(\chi_{4-j}^2 \geq \lambda_j^4(x)\right) > \alpha;\ (2 \leq j \leq 3)$, or until no more comparisons can be performed, in which case, the best tree from group Θ_4 is accepted (see Fig. 7A).

Stepwise method: We start by calculating the likelihood ratio $\lambda_3^4(x) = -2\log\frac{L(x|\Theta_3)}{L(x|\Theta_4)}$, and the P-value $p_3^4(x) \approx \mathrm{P}\left(\chi_1^2 \geq \lambda_3^4(x)\right)$. If $p_1^4(x) < \alpha$, the best tree of group θ4 is accepted; if not we proceed to test the best tree from group θ2 against the best tree from group θ3. We repeat these comparisons until $p_j^{j+1}(x) \approx \mathrm{P}\left(\chi_1^2 \geq \lambda_j^{j+1}(x)\right) < \alpha\ (j = 2, 1, 0)$, or until no more comparisons can be performed, in which case, polytomy is assumed (Fig. 7B).

Given the complexity of the calculations and the optimization formulas for maximizing the likelihood function $l(x;\theta)$, we initially used Wolfram Mathematica 10 (https://www.wolfram.com/mathematica/) to check the source formulas and to test various optimization options.

Simulations. To test the likelihood ratio values against the χ_k^2 distribution, we carried out various simulations. We denoted $\lambda_j^k(\xi)$ to be a random variable whose distribution depends on θ. We then fixed *j* and *k*, as well as $\theta \in \Theta_j$, and defined $\mathrm{Q}(z) = \mathrm{P}_\theta\left(\lambda_j^k(\xi) \geq z\right)$. After running the simulation more than 1000 times, we obtained different sets of ξ values, and for

each of them we calculated the values $L(\xi|\Theta_k)$ and $L\left(\xi|\Theta_j\right)$, and then used them to calculate $\lambda_j^k(\xi)$. According to the *Law of Large Numbers* (e.g., see in Grimmett and Stirzaker, 1992), replacing probability by frequency, we can write approximately: $Q(z) \approx \frac{m(z)}{n}$, where *m* (*z*) – the number of values $\lambda_j^k(\xi)$ – is more than or equal to z.

4. Testing phylogenetic diagnostic markers

To demonstrate the effectiveness of the statistics, we used them to determine the significance of retroposon data supporting known phylogenetic trees of primates, domestic dogs, and mouse strains. We used the 2-n-way suite to generate 2-way genome alignments and to extract orthologous presence/absence retrotransposon loci in fasta format. Manual inspection of orthology and correction of MUSCLE alignments ensured reliable information for the computationally extracted loci. Such verified loci were then used to derive the frequency of diagnostic retrotransposon insertions sorted by the ten possible tree topologies for four species.

Great ape phylogenetic project

To investigate the phylogenetic significance of presence/absence patterns of active SINE-VNTR-Alu SINEs in great apes, we screened all available great ape genomes. Human (*Homo sapiens*, version hg38), chimpanzee (*Pan troglodytes*, version Clint_PTRv2), bonobo (*Pan paniscus*, version panPan1.1), gorilla (*Gorilla gorilla*, version gorGor4), and orangutan (*Pongo abelii*, version Susie_PABv2) genomes (see Appendix 3) were used to generate 2-way genome alignments (Churakov et al., 2020) (http://retrogenomics.uni-muenster.de/tools/twoway) with human as target genome (human-chimpanzee, human-bonobo, human-gorilla, human-orangutan). With n-way (Churakov et al., 2020)

http://retrogenomics.uni-muenster.de/tools/nway), we then created two sets of four species: human, chimpanzee, gorilla, orangutan, and human, chimpanzee, bonobo, gorilla. To extract presence/absence patterns for the ten possible tree topologies for four species, we ran one direct (starting from the human target) and three reverse searches (starting from the three different query genomes) in n-way. We restricted our search to diagnostic SINE-VNTR-*Alu* (SVA) retrotransposons that were active during the diversification of great apes. A maximum of 10-nt truncated ends of elements were allowed, all duplications were removed, and only perfect matches were considered.

Dog lineage diversification project

To sort the phylogenetic history of domestic dog breeds we used boxer (version CanFam3.1, GCA_000002285.2), beagle (version Beagle, GCA_000331495.1), dingo (version ASM325472v1, GCA_003254725.1), and German shepherd (version ASM864105v1, GCA_008641055.1) (see Appendix 3). We built an n-way project for these four dog breeds, with boxer as the target species and the remaining species as queries. All genomes and RepeatMasker reports were downloaded from NCBI. The fastCOEX tool (http://retrogenomics.uni-muenster.de/tools/fast COEX/index.hbi?; (Doronina et al., 2017b) was applied to extract nearly full-length (<6-nt truncations) genomic dog-specific SINE elements (SINEC_Cf, SINEC_Cf2, and SINECA1_Cf), flanked by repeat-free sequences. We performed one direct and three reverse n-way searches with coordinates of the selected SINEs. All duplicates were removed, and the resulting perfect presence/absence patterns were downloaded as aligned fasta files.

Mouse strains project. To confirm the phylogenetic relationships among inbred laboratory mouse strains, we download their genomes and RepeatMasker reports from NCBI (see Appendix 3). We utilized CBA/J (CBA_J_v1), C57BL6/J (ASM377452v2), BALB/cJ (BALB_cJ_v1), and DBA/2J (DBA_2J_v1) mouse strain genomes to generate six 2-way genome alignments with CBA/J or C57Bl6/j selected as targets (http://retrogenomics.uni-muenster.de/tools/twoway). We build an n-way project for these four mouse strains and performed two direct n-way searches for SINE/*Alu* and SINE/B2 elements from CBA/J and C57Bl6/J strains (http://retrogenomics.uni-muenster.de/tools/nway). Perfect presence/absence patterns with all duplications removed were downloaded as aligned fasta files.

Results and Discussion

Four-lineage phylogenies require the evaluation of 10 predefined presence/absence patterns for each inserted retroposon. In developing these statistics, we introduced y-based short names to describe the individual patterns with the two indices *i* and *j*. *i=j* denotes one absence and three presences (e.g., y_{11} = -+++). *j>i* denotes two absences and two presences (e.g., y_{23}= +--+). We then sorted all ten presence/absence patterns in ascending order as follows: 1) [y_{11}] -+++; 2) [y_{12}] --++; 3) [y_{13}] -+-+; 4) [y_{14}] -++-; 5) [y_{22}] +-++; 6) [y_{23}] +--+; 7) [y_{24}] +-+-; 8) [y_{33}] ++-+; 9) [y_{34}] ++--; 10) [y_{44}] +++- (note: neither ++++ nor patterns with 3 minuses are phylogenetically informative).

We built two statistical algorithms with varying sensitivities to potentially short branches. The more robust and routinely used *reverse* algorithm was less reliable in detecting critical short branches, and correspondingly, in handling small numbers of phylogenetic markers.

The *stepwise* algorithm was adapted to detect partly resolved trees. For example, the distribution y_{11}=22; y_{12}=25; y_{13}=7; y_{14}=11; y_{22}=14; y_{23}=12; y_{24}=18; y_{33}=16; y_{34}=17; y_{44}=24 was unresolved by the reverse algorithm, while the stepwise algorithm revealed a partially resolved solution (PT model; H1:0Tn1:1234, T_3=0.13).

Command-line likelihood estimator

We developed a standalone python console script (Python 2.7/3.6 or higher) we call hammlet (Hybridization Models Maximum Likelihood Estimator) that can be installed on various operating systems. Hammlet performs likelihood calculations for different tree topologies, utilizes two different methods of statistical comparisons, finds essential trees for the data to be evaluated, and draws these trees. The command-line procedure leads the user from their data to the final tree with statistical evaluation and requires input of the frequencies of phylogenetic markers supporting each of the ten possible, carefully evaluated presence/absence patterns (see above) as well as some additional parameters. It should be noted that the order of values in the user data vector is predetermined (see above). A typical result provides the user with the following sorts of information: the level or group (N0: θ0, polytomy; N1: θ1, one parameter released, three fixed; N2: θ2, two parameters released, two fixed; N3: θ3, three parameters released, one fixed; and N4: θ4, four parameters released; the model (denoted Tx), the model-variant and its alias name (e.g., H1:TT01), the permutation of the species orders (e.g., 1324), the likelihood (LL), the effective population size coefficient (n0), the length of the first branch (T1), the length of the second branch (T3), the gamma value for the first hybridization (g1), and the gamma value for the second hybridization (g3)). The next step is the drawing of the maximum likelihood tree using the command-line: *hammlet draw*, followed by the information provided by the above

parameters, wherein the initial order of species names A,B,C,D is replaced by A,C,B,D for the permutation 1324. The derived tree outfile is in svg-format and can then be visualized in any internet browser.

In addition to the procedure just described, the command-line interface can also calculate optimized likelihood values for user-defined sets of models and permutations but without statistical calculations (output in a user-defined file with comma-delimited format). For developers, the command-line interface provides a reverse technique to derive specific marker distributions from a preset phylogenetic tree topology. Here, the number of ten $\gamma_{i,j}$ marker compositions are derived from the five optimized parameters (n0, T1, T3, g1, g3) with a subsequent bootstrap optimization step. This can be used for simulation of marker-tree variations. The command-line script is available upon request.

R-based web interface

We used the R-programming language to develop a user-friendly interface. The latest version of hammlet is integrated into a shiny server. After the user has input the observed frequencies of phylogenetic markers for all of the ten predefined presence/absence patterns for four lineages, they then select the statistical method to be used (reverse or stepwise) and significance (cut-off) values. The program then calls hammlet to optimize parameters and calculate the likelihood of the 155 hypothetical trees, including unique and multiple hybridization scenarios as well as polytomies. For each of the five predefined groups (θ0-θ4, see Methods), the tree with the best likelihood value is selected and statistically compared to the next likely tree. It then provides a visualization of the tree from the best-fitted model and a table of detailed parameters used to find the tree with the highest likelihood. The

statistical difference between the best model and polytomy is then shown under the tree figure. For the tree determined to be the most likely, an implemented KKSC module calculates the significance for all individual nodes (splits/hybridizations) of the tree presented in the middle section of the final screen. A table of the tree parameters is shown at the bottom section of the screen (see instructions on the application page http://retrogenomics.uni-muenster.de:3838/hammlet/instructions.html). The figure of the likeliest tree and the table of detailed parameters can be downloaded. It is also possible to upload a file containing the algorithms that were applied, the significance cutoff values chosen, and the frequencies of phylogenetic markers inputted.

Simulation

The chi-square test was applied to test the statistical significance of different likelihood ratios for the different tree topologies. Here we used simulated presence/absence frequencies to compare their chi-square distributions. For each set of $\theta = (\mathrm{m}, \sigma, \nu)$ we selected values of model type (2H1 or 2H2, restricting permutation to "1234") and parameter values $(n_0, T_1, T_3, \gamma_1, \gamma_3)$ corresponding to the selected tree Θ_j. Then, using specific formulas (see Appendix 1, S1.3.1.-S1.3.30), we calculated $a_{i,j}$ and, finally, using a pseudo-random number generator (created in the Wolfram Mathematica 10; https://www.wolfram.com/mathematica/), we evaluated ten independent random variables $\xi_{i,j}$ dispersed according to the Poisson distribution with parameters $a_{i,j}$. Using the hammlet program, for each set of $\xi_{i,j}$, we then found the optimized likelihood values $L(\xi|\Theta_j)$ and $L(\xi|\Theta_k)$, from which we calculated $\lambda_j^k(\xi)$. For both the reverse (stringent) and stepwise (relaxed) statistical methods, for each case, we set the appropriate values of free parameters and considered the corresponding distribution $\lambda_j^k(k = 4)$. Then

we created comparative distribution graphs as presented in Appendix 2, TableS1, which presents a mean, standard deviation between the empirical distributions, and a chi-square approximation of about 4.3% at a significance level α=0.05. The distribution curves showed similar sigmoid distributions. Corresponding to these data, a good agreement was found between $Q(z)$ and $\mathrm{P}\left(\chi^2_{k-\mathrm{j}} \geq \mathrm{z}\right)$.

Connection between trees and presence/absence data patterns

The 4-LIN test can easily designate significantly supported phylogenetic trees based on conflict-free inserted phylogenetic presence/absence data without maximum likelihood optimization. However, in cases when conflicting markers are present that interfere with the reconstruction of simple trees, one of two variants of partial polytomy emerges; for example, a non-zero value for y_{12} can be interpreted as a PT tree (H1:0Tn1:1234), and a non-zero value for y_{44} can be construed as a TP tree (H1:T00n:4123). Two non-zero values in some cases can also be interpreted as a completely non-conflicting tree; for example, non-zero values for y_{13} and y_{24} can be construed as a T1 tree (H1:TT10:1234), and finally, non-zero values for y_{33} and y_{34} can be interpreted as a T2 tree (H1:TT01:3412). It should be mentioned that in the case of two non-zero values and the T1 model, the order of lineages depends on the ratio of the value; for example, in the case of y_{13} and y_{24}, if $y_{13}>y_{24}$ the order of species will be 1234, while in the opposite case, it will be 2143. However, the T2 model is free of such conditions. In Appendix 2, TableS2, all possible tree variants defined by one or two non-zero values are presented and can be used as a reference list for non-conflicting datasets.

To derive a simple phylogenetic tree from the often diverse and sometimes conflicting presence/absence patterns of phylogenetic markers, taking in account stochastic variation and different possible permutations, it is necessary to use statistical applications. However, KKSC or other statistical applications, which take into consideration only qualitative criteria, are less efficient at finding the actual tree topology for four lineages. In contrast to these strict qualitative approaches, the 4-LIN *quantitative* approach, based on the maximum likelihood ratio, evaluates each possible branch independently and summarizes this information only at the last stage of likelihood calculation to efficiently reconstruct the complete 4-LIN tree.

However, the significance values calculated by the hammlet script refer to support for the entire tree topology and not the individual branches of the tree. Individual significance values for specific branches are more complex to derive because the fixation of a marker at one branch is independent of fixation at neighboring branches (Kuritzin et al., 2016). To also obtain support-values for individual branches, we embedded elements of the KKSC statistical calculation (Kuritzin et al., 2016) to compute the significance of each split. Using the KKSC method in combination with the 4-LIN statistics to designate specific support for individual branches requires six values (e.g., for tree T1 and order of species 1234 in a first comparison $Y_1=y_{24}+y_{44}$; $Y_2=y_{23}+y_{33}$; $Y_3=y_{12}+y_{11}$; and in a second comparison $Y_1=y_{13}+y_{33}$; $Y_2=y_{12}+y_{22}$; $Y_3=y_{14}+y_{44}$). In Appendix 2, TableS3 for all possible resolved trees, we present formulas to connect the results of the 4-LIN test with the KKSC statistics.

Examples for well-known phylogenetic relations

Apes: SINE elements were already used as phylogenetic markers to resolve the great ape sister relationship between humans and chimpanzees (Salem et al., 2003). To test the 4-LIN statistics in great apes, we first searched the genomes of human, chimpanzee, gorilla, and orangutan for SVA SINE elements. This screen detected 56 diagnostic markers whose presence/absence patterns were distributed as follows: y_{11}=0; y_{12}=0; y_{13}=0; y_{14}=0; y_{22}=0; y_{23}=0; y_{24}=1; y_{33}=0; y_{34}=37; y_{44}=18. Running the 4-LIN statistical test for this dataset (with both the stringent reverse and relaxed stepwise algorithms) and a cut-off value of P = 0.05, we found the maximum likelihood for tree T2 (H1:TT01:4312, T_1=1.93, T_3=3.36, $p < 2.06 \cdot 10^{-09}$), which confirms the current view on hominid phylogeny (((human, chimpanzee), gorilla), orangutan; (Salem et al., 2003) (Fig. 8A). Here both individual splits were also shown to be significantly supported (p = 2.58e-09, p = 5.7e-17, KKSC test).

We also screened a second set of species comprising human, chimpanzee, bonobo, and gorilla. The 52 detected presence/absence patterns revealed the following distribution: y_{11}=0; y_{12}=0; y_{13}=0; y_{14}=43; y_{22}=0; y_{23}=1; y_{24}=0; y_{33}=0; y_{34}=3; y_{44}=5. The 4-LIN test for this species set favored the hybridization model 1H3 (H1:TT0g:4132, T_1=0.81, T_3=8.01, γ_3=0.94). The most likely supported tree shows a close relationship of chimpanzee and bonobo with humans as the sister group and gorilla as the outermost diversification (however, the KKSC test did not show significant support for this split (p = 0.33)). This also agrees with the current view of their relationships (Salem et al., 2003). However, we also detected a distinct signal of ancestral hybridization/introgression (6%) between human and chimpanzee (this hybridization/introgression was not detected by the KKSC test; p = 0.25; Fig. 8B).

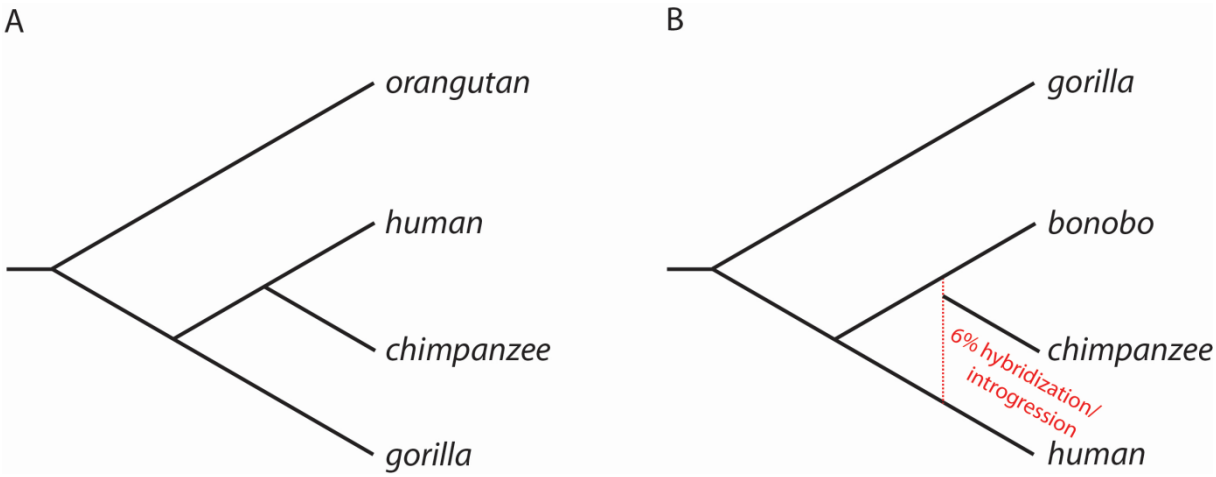

Figure 8. Phylogenetic trees of hominids. Black lines indicate tree branches; the red vertical dotted line shows hybridization/introgression.

Dogs: Domestic dogs (*Canis lupus familiaris*) evolved from the wolf lineage about 20-40 thousand years ago and have since been diversified into approximately 400 different breeds (Galibert et al., 2011). The large number of dog-specific SINE elements that were active over a long period of time (Peng et al., 2018) and the rapid diversification of lineages boosted by domestication makes domesticated dogs a good example for testing the 4-LIN statistical approach. We first used 2-n-way (Churakov et al., 2020) to find and extract diagnostic presence/absence markers from 2-way alignments generated from the genomes of the boxer, beagle, German shepherd, and dingo. This screen yielded 1534 SINE markers distributed over all 10 predefined presence/absence patterns as follows: y_{11}=144; y_{12}=287; y_{13}=148; y_{14}=87; y_{22}=533; y_{23}=370; y_{24}=288; y_{33}=345; y_{34}=161; y_{44}=143. The initial order of breeds was boxer, beagle, German shepherd, dingo. After running the 4-LIN test using both the reverse and stepwise algorithms and a cut-off value of p = 0.05, we found significant support for the model 1H4 ((H1:TTg1:2314) T_1=0.79; T_3=0.28; γ_1=0.35, $p<1 \cdot 10^{-64}$), where boxer and shepherd are closest (p = 3.27e-23, KKSC test), the dingo is the sister group to them, and beagle is the most distant. However, the beagle, which was used as an outgroup,

showed 35% hybridization/introgression signals with the boxer/German shepherd ancestral branch (p = 6.54e-23, KKSC test; Fig. 9).

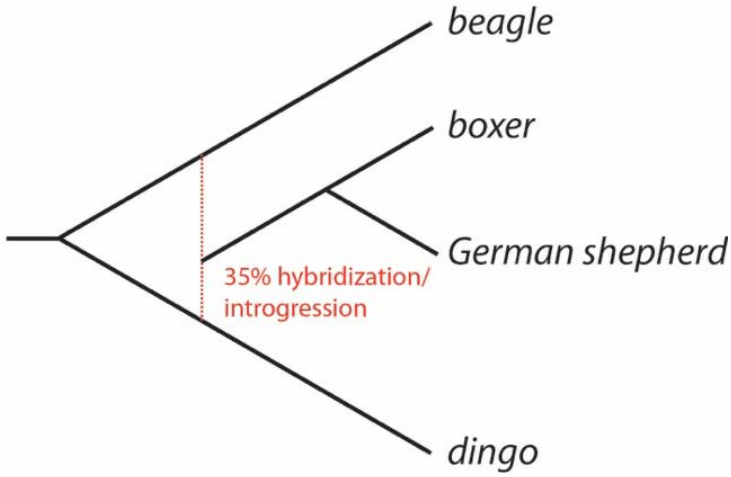

Figure 9. Phylogenetic tree of domestic dog breeds. Black lines indicate tree branches; red vertical dotted line shows hybridization/introgression events

The boxer and German shepherd fusion as well as their relationship to dingo confirm the findings of other phylogenetic studies (Parker, 2012; Parker et al., 2017). Interestingly, the hybridization/introgression signal of beagle may have arisen with the early diversification of boxer/German shepherd. We expect that an indirect hybridization/introgression via crossing with wolves might have been a substantial source of such incidences (vonHoldt et al., 2012). Alternatively, hybridization is also ordinary for such artificial selections and well known for dogs (vonHoldt et al., 2010; Wayne and vonHoldt, 2012). Our dog breed results show that the 4-LIN test can be helpful not only to reconstruct deep phylogenetic events but also to evaluate the history of domesticated breeds.

Mouse inbred strains: Mouse inbred strains started their human-laboratory diversification around 175 years ago (Atchley and Fitch, 1991). In the meantime, more than 110 stable strains are distributed in scientific laboratories across the world. Due to their inbred nature these strains carry many metabolic changes and are used in a wide range of biological experiments (Ghazalpour et al., 2012). For our investigations, we selected four mouse inbred strains: CBA/J, C57BL6/J, BALB/cJ, and DBA/2J located most distantly on the phylogenetic mouse tree. A screen for SINE/*Alu* and SINE/B2 elements yielded 2575 markers distributed over all 10 predefined presence/absence patterns as follows: y_{11}=341; y_{12}=185; y_{13}=215; y_{14}=157; y_{22}=356; y_{23}=197; y_{24}=258; y_{33}=197; y_{34}=147; y_{44}=522. The 4-LIN test run with the reverse algorithm and a cut-off value of p = 0.05, showed significant support for the 1H4 model (H1:TTg1:2413, T_1=0.22; T_3=0.17; γ_1=0.7, $p<1 \cdot 10^{-64}$, Fig. 10). The tree confirms a consolidation of CBA/J and BALV/cJ strains against C57Bl/6J and DBA/2J strains ($p=2.18e^{-12}$, KKSC). However, a hybridization/introgression scenario in the ancestry of the C57Bl/6J and DBA/2J strains is also significantly supported (p = 0.00003, KKSC test). Such an ancestral hybridization/introgression event might indicate the documented crosses between strains during the early stages of modern mice laboratory diversification (Atchley and Fitch, 1993; Atchley and Fitch, 1991).

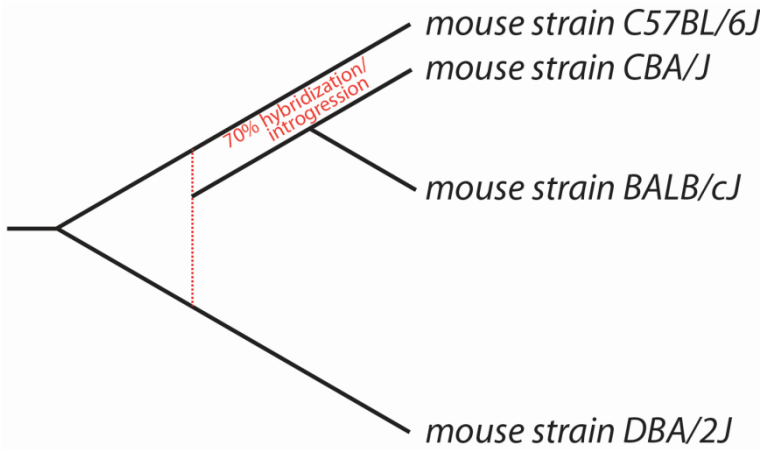

Figure 10. Phylogenetic tree of four inbred mouse strains. Black lines indicate the tree branches; the red vertical dotted line shows the ancestral hybridization/introgression event.

Direct comparison of significance levels to KKSC

The KKSC method to evaluate the significance of retroposon insertion data support for phylogenetic trees that contain evidence of ILS and hybridization requires three-lineage combinations. Given a situation where three diagnostic markers all support the first of these three investigated tree topologies, according to Kuritzin et al. (2016), a pattern of 3:0:0 indicates an initial significance value of $p<0.05$. If we provide a comparable 4-LIN vector of 10 $y_{i,j}$ values where all $y_{i,j}$ are 0, except $y_{1,3} = 3$ (e.g., vector: 0, 0, 3, 0, 0, 0, 0, 0, 0, 0), and select the lineage order A,B,C,D, the 4-LIN test yields significant support for the consolidation of the branches B and D, where A, C, and the ancestral branch of B and D form a polytomy ($p=0.01022$ for the overall tree topology). To obtain significant support for the two splits indicated by the 4-LIN software we need at least four markers supporting the two splits (e.g., vector: 4, 0, 4, 0, 0, 0, 0, 0, 0, 0). This marker distribution will generate a significant tree (A,(C,(B,D))) with $p=0.01590$.

Conclusions

Based on the ten possible presence/absence distribution patterns of phylogenetically informative markers for four lineages, the 4-LIN test calculates the phylogenomic tree with the maximum likelihood. It determines this tree's statistical significance compared to the next less-supported candidate tree. The need to provide a 4-LIN test and to develop the complex underlying mathematical framework arose from the growing interest in more

extensive phylogenomic comparisons of whole-genome retrotransposon presence/absence patterns, where multiple comparisons of groups of only 3 species was inefficient and often ineffective. Compared to the currently available KKSC statistics for evaluating phylogenomic data that compares only seven tree variants, the 4-LIN test is finetuned to compare 155 different trees, including scenarios for multiple ancestral hybridizations/introgressions and incomplete lineage sorting. The contribution of ancestral hybridization/introgression to the phylogeny of these species can now be quantified and applied to the reconstruction of the complete tree topology. Moreover, the 4-LIN statistical approach can be combined with the KKSC statistics to thoroughly evaluate complete tree topologies including their individual branches. However, the 4-LIN test is in general more sensitive than the KKSC and can resolve weak ancestral hybridization/introgression signals or very short branches (see e.g., second hominid example). While the 4-LIN test is sufficient for resolving most phylogenomic conundrums, the ultimate goal of developing the 4-LIN test was to derive a strategy for an n-lineage analysis operating on complete presence/absence matrices of any number of lineages. The maximum likelihood approach and all necessary mathematical background calculations we have described here provide the necessary technical underpinnings to target such a goal. The current advantage of 4-LIN comparisons is that the results of extended screenings involving four instead of three species are directly applicable to resolve more complex phylogenetic questions. This enables evaluation of data collected for ten different potential elementary tree topologies instead of just three with KKSC. As examples, we used the 4-LIN test to evaluate the significance of phylogenetic marker distributions in great apes, dog breeds, and mouse strains.

Software availability

The new Web tool, including the methods presented here, can be found at:

http://retrogenomics.uni-muenster.de:3838/hammlet/

Acknowledgments

We thank Marsha Bundman for editorial help. We thank Norbert Grundmann for his help in integrating the shiny framework into a web-server. This work was supported by the Deutsche Forschungsgemeinschaft (DFG) grant number (SCHM 1469/10-1 to JS) and DFG grant number (281125614/GRK2220 to the Research Training Group Evolutionary processes in Adaptation and Disease [EvoPAD]).

Author contributions

GC, AK, and JS conceived the 4-lineage project. GC and AK developed and optimized the statistical strategy. CK and VU designed python scripts for automatic likelihood maximization and statistical evaluation. FZ and FW designed R-scripts for the web interface and statistical evaluation. GC and JS wrote the paper with input from all other authors.

Competing interests

The authors declare no competing interests.

References

Atchley, W. R., and W. Fitch. 1993. Genetic affinities of inbred mouse strains of uncertain origin. Mol Biol Evol 10:1150-69.

Atchley, W. R., and W. M. Fitch. 1991. Gene trees and the origins of inbred strains of mice. Science 254:554-8.

Churakov, G., F. Zhang, N. Grundmann, W. Makalowski, A. Noll, L. Doronina, and J. Schmitz. 2020. The multicomparative 2-n-way genome suite. Genome Res 30:1508-1516.

Doronina, L., G. Churakov, A. Kuritzin, J. Shi, R. Baertsch, H. Clawson, and J. Schmitz. 2017a. Speciation network in Laurasiatheria: retrophylogenomic signals. Genome Res 27:997-1003.

Doronina, L., A. Matzke, G. Churakov, M. Stoll, A. Huge, and J. Schmitz. 2017b. The Beaver's Phylogenetic Lineage Illuminated by Retroposon Reads. Sci Rep 7:43562.

Doronina, L., O. Reising, H. Clawson, D. A. Ray, and J. Schmitz. 2019. True Homoplasy of Retrotransposon Insertions in Primates. Syst Biol 68:482-493.

Galibert, F., P. Quignon, C. Hitte, and C. Andre. 2011. Toward understanding dog evolutionary and domestication history. C R Biol 334:190-6.

Ghazalpour, A., C. D. Rau, C. R. Farber, B. J. Bennett, L. D. Orozco, A. van Nas, C. Pan, H. Allayee, S. W. Beaven, M. Civelek, R. C. Davis, T. A. Drake, R. A. Friedman, N. Furlotte, S. T. Hui, J. D. Jentsch, E. Kostem, H. M. Kang, E. Y. Kang, J. W. Joo, V. A. Korshunov, R. E. Laughlin, L. J. Martin, J. D. Ohmen, B. W. Parks, M. Pellegrini, K. Reue, D. J. Smith, S. Tetradis, J. Wang, Y. Wang, J. N. Weiss, T. Kirchgessner, P. S. Gargalovic, E. Eskin, A. J. Lusis, and R. C. LeBoeuf. 2012. Hybrid mouse diversity panel: a panel of inbred mouse strains suitable for analysis of complex genetic traits. Mamm Genome 23:680-92.

Grimmett, G. R., and D. R. Stirzaker. 1992. Probability and Random Processes: Problems and Solutions, 3rd edition. Oxford University Press.

Kimura, M. 1955a. Solution of a Process of Random Genetic Drift with a Continuous Model. Proc Natl Acad Sci U S A 41:144-50.

Kimura, M. 1955b. Stochastic processes and distribution of gene frequencies under natural selection. Cold Spring Harb Symp Quant Biol 20:33-53.

Kuritzin, A., T. Kischka, J. Schmitz, and G. Churakov. 2016. Incomplete Lineage Sorting and Hybridization Statistics for Large-Scale Retroposon Insertion Data. PLoS Comput Biol 12:e1004812.

Parker, H. G. 2012. Genomic analyses of modern dog breeds. Mamm Genome 23:19-27.

Parker, H. G., D. L. Dreger, M. Rimbault, B. W. Davis, A. B. Mullen, G. Carpintero-Ramirez, and E. A. Ostrander. 2017. Genomic Analyses Reveal the Influence of Geographic Origin, Migration, and Hybridization on Modern Dog Breed Development. Cell Rep 19:697-708.

Peng, C., L. Niu, J. Deng, J. Yu, X. Zhang, C. Zhou, J. Xing, and J. Li. 2018. Can-SINE dynamics in the giant panda and three other Caniformia genomes. Mob DNA 9:32.

Salem, A. H., D. A. Ray, J. Xing, P. A. Callinan, J. S. Myers, D. J. Hedges, R. K. Garber, D. J. Witherspoon, L. B. Jorde, and M. A. Batzer. 2003. Alu elements and hominid phylogenetics. Proc Natl Acad Sci U S A 100:12787-91.

vonHoldt, B. M., J. P. Pollinger, D. A. Earl, H. G. Parker, E. A. Ostrander, and R. K. Wayne. 2012. Identification of recent hybridization between gray wolves and domesticated dogs by SNP genotyping. Mamm Genome 24:80-8.

vonHoldt, B. M., J. P. Pollinger, K. E. Lohmueller, E. Han, H. G. Parker, P. Quignon, J. D. Degenhardt, A. R. Boyko, D. A. Earl, A. Auton, A. Reynolds, K. Bryc, A. Brisbin, J. C.

Knowles, D. S. Mosher, T. C. Spady, A. Elkahloun, E. Geffen, M. Pilot, W. Jedrzejewski, C. Greco, E. Randi, D. Bannasch, A. Wilton, J. Shearman, M. Musiani, M. Cargill, P. G. Jones, Z. Qian, W. Huang, Z. L. Ding, Y. P. Zhang, C. D. Bustamante, E. A. Ostrander, J. Novembre, and R. K. Wayne. 2010. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature 464:898-902.

Waddell, P. J., H. Kishino, and R. Ota. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Inform 12:141-54.

Wayne, R. K., and B. M. vonHoldt. 2012. Evolutionary genomics of dog domestication. Mamm Genome 23:3-18.