II Articles

Gessica Sakamoto Martini, Jeremy Kendal, and
Jamshid Johari Tehrani

Cinderella's Family Tree. A Phylomemetic Case Study of ATU 510/511

https://doi.org/10.1515/fabula-2023-0002

Abstract: This case study contributes to recent attempts to apply "phylomemetic" methods derived from computational biology to oral traditions, where the aim is to trace the mutation and diversification of folk narratives as they get passed on from generation to generation and spread from society to society. Our study focuses on one of the most famous and widespread tales in the folktale record: Cinderella. Thousands of Cinderella-like stories have been documented from around the world, which folklorists have attempted to classify into different "types" representing distinct, though related, international traditions. The most comprehensive of Cinderella typologies was developed by Anna Birgitta Rooth (1951), who divided the tales into five principal types: A, B, AB, BI and C, and suggested several hypotheses pertaining to their origins and relationships to one another. Here, we test Rooth's theories on a sample of 266 versions of Cinderella using Bayesian phylogenetic inference, phylogenetic networks (NeighborNet) and a model-based clustering method that was originally designed to elicit population structure from multi-locus genotype data (implemented in the program STRUCTURE). Our results find varying levels of support for the types identified by Rooth, and suggest that mixing among traditions was widespread, especially in Type AB tales. Despite these complexities, it was still possible to delineate and quantify the influence of distinct ancestral sources on the variation observed in contemporary versions of Cinderella. Our study highlights the value and versatility of phylomemetic methods in uncovering the historical relationships among types and sub-types of international folktale, as well as the evolutionary processes that have shaped them.

Gessica Sakamoto Martini, Durham Cultural Evolution Research Centre, Durham University, Durham, U.K. E-mail: gessicamartini90@gmail.com
Jeremy Kendel, Department of Anthropology, Durham Cultural Evolution Research Centre, Durham Research Methods Centre, Durham University, Durham, U.K. E-mail: jeremy.kendal@durham.ac.uk
Jamshid Johari Tehrani, Durham Cultural Evolution Research Centre, Department of Anthropology, Durham University, Durham, U.K. E-mail: jamie.tehrani@durham.ac.uk

Zusammenfassung: Diese Fallstudie versteht sich als Beitrag zu den Versuchen der letzten Jahre, aus der Bioinformatik stammende „phylomemetische“ Methoden auf mündliche Überlieferungen anzuwenden. Ziel dieser Ansätze ist es, Wandel und Ausdifferenzierung der Erzählungen in der Weitergabe von einer Generation zur anderen sowie im Übergang von einer Gesellschaft zur anderen nachzuzeichnen. Unser Beitrag konzentriert sich auf einen der bedeutendsten und weitverbreitetsten Märchenstoffe überhaupt: Aschenputtel. Tausende von Aschenputtel-artigen Erzählungen aus der ganzen Welt sind bekannt. Erzählforscher haben versucht, diese verschiedenen „Typen“ zuzuordnen, welche unterschiedliche, allerdings miteinander zusammenhängende internationale Überlieferungen darstellen. Die umfassendste Aschenputtel-Typologie wurde von Anna Birgitta Rooth (1951) entwickelt. Sie unterscheidet fünf Haupttypen (A, B, AB, BI und C) und hat verschiedene Thesen zu deren Ursprüngen und Beziehungen zueinander vorgelegt. Wir überprüfen hier die Rooth'schen Thesen anhand einer Stichprobe von 266 Aschenputtel-Erzählungen, indem wir bayessche phylogenetische Inferenz, phylogenetische Netzwerke (NeighborNet) sowie (durch Einsatz der Software „STRUCTURE“) eine modellbasierte Clusterbildungsmethode einsetzen. Letztere wurde ursprünglich entwickelt, um anhand von Multi-Lokus-Genotyp-Daten auf Populationsstrukturen zu schließen. Unsere Ergebnisse bestätigen die von Rooth festgestellten Typen in unterschiedlichem Maß und legen nahe, dass – besonders bei den Erzählungen des Typs AB – eine Vermischung der Überlieferungsstränge den Normalfall darstellte. Trotz dieser Schwierigkeiten konnten wir dennoch den Einfluss bestimmter Vorläufer auf die in zeitgenössischen Aschenputtel-Versionen zu konstatierende Variation skizzieren und quantitativ bestimmen. Unser Beitrag veranschaulicht den Wert und die Vielseitigkeit phylomemetischer Methoden bei der Sichtbarmachung der historischen Beziehungen zwischen Typen und Untertypen des internationalen Märchens sowie der Entwicklungsvorgänge, die sie geprägt haben.

1 Introduction

The construction of narrative typologies has long been central to efforts to understand cross-cultural relationships in traditional stories. In the late nineteenth and twentieth centuries, proponents of the “historic-geographic” school sought to classify folktales, legends and jokes into distinct “tale types” based on a core set of shared “motifs” (characters, artefacts or episodes) that are highly stable in their transmission (Aarne/Thompson 1961; Thompson 1977; Goldberg 1984). By assembling all the known variants of a given international type and sorting them by region and chronology, these researchers sought to locate the sources and home-

lands of common folktales, track their routes of diffusion, and reconstruct their original *ur*-forms. These efforts were often explicitly inspired by the role played by biological taxonomies in studying the evolutionary relationships among species, as exemplified by Stith Thompson's comment that

> biologists have long since labelled their flora and fauna by a universal system and by using this method have published thousands of inventories of the animal and plant life of all parts of the world [...] The need for such an arrangement of narrative has been realized for a long time. (Thompson 1977, 414)

While the historic-geographic school can boast many important and lasting achievements in cataloguing and reconstructing cross-cultural relationships among folktales (e.g., Uther 2004), it did not produce a fully-fledged taxonomy of folklore equivalent to modern biological systematics. As critics of the approach have pointed out, this is because tale types are typically based on just a few motifs that are often highly ethnocentric and difficult to apply in wider comparative contexts (e.g., Dundes 1997; Goldberg 1984). Recently, however, some researchers have sought to address these limitations by developing more systematic, quantitative approaches that draw on modern computational methods from phylogenetics and population genetics (Tehrani/d'Huy 2017). Originally developed to study genetic relationships within and between species, these techniques have become increasingly adopted in other fields, including historical linguistics, archaeology, and textual analysis (Howe/Windram 2011). These applications of phylogenetics and population genetics have been labelled as "phylomemetics" (Howe/Windram 2011), since they focus on the transmission of cultural information, or "memes"[1] (after Dawkins 1976), rather than genes. The key objective for both phylogenetic and phylomemetic analyses is to reconstruct historical relationships among a group of entities by excavating information about the past that has been preserved through the mechanism of inheritance. In biology, this information typically consists of mutations in sequences of DNA, whereas in cultural data it may comprise changes in word forms, cumulative innovations in a technological or craft tradition, or scribal errors found in texts copied from the same exemplar (Howe/Windram 2011). In the case of folktales, mutations and adaptations resulting from the repeated re-telling of a story across generations (e.g., the gender of the protagonists, types of animals and supernatural characters, changes to the ending of the tale, and so on) can be

1 "Meme" was coined by Richard Dawkins (1976), but it is worth noting that phylomemetics does not entail any theoretical commitment to Dawkins' wider proposals concerning the parallels between genes and memes (such as both operating as "selfish" replicators). We use the term in a more restricted sense to refer to units of cultural inheritance.

used to discriminate distinct lineages of transmission, and model their relationships to one another (Tehrani/d'Huy 2017). A major advantage of this approach is that it takes into account all the resemblances among a set of tales, rather than basing taxonomic groups on a few privileged, pre-determined motifs.

While research in this area is still in its infancy, phylomemetic studies of folktales have demonstrated that it is possible to reconstruct the transmission histories of several tale types, including *ATU 480 The Kind and Unkind Girl* (Ross et al. 2013), *ATU 1137 Polyphemus* (d'Huy 2015), Pygmalion (d'Huy 2013), *ATU 333 Little Red Riding Hood* and *ATU 123 the Wolf and the Kids* (Tehrani 2013). The latter study is especially relevant in the current context since it specifically addressed on how phylomemetic methods can help to resolve some of the ambiguities inherent in traditional tale taxonomies. It has long been known that the distinction between *ATU 333* and *ATU 123* is problematic: both tales concern a dangerous predator (usually a wolf) who attacks their victim(s) by posing as a relative. In European and Middle Eastern traditions, these types are differentiated by two key features: whether the victims of the tale are human (*ATU 333*) or animal (*ATU 123*), and whether the predator (usually a wolf) attacks them in their own home (*ATU 123*) or at their grandmother's house (*ATU 333*). However, there are highly similar and clearly related tales in parts of Africa and East Asia which defy this categorisation. In many of these tales, the victim is human (like *Little Red Riding Hood*) but they are attacked in their own home (as in *The Wolf and the Kids*). By taking a quantitative, phylomemetic approach that incorporated a much wider range of traits (72 in total), Tehrani was able to establish that the African tales clearly group with *ATU 123 The Wolf and the Kids*. The East Asian stories, meanwhile, formed a separate lineage distinct from both *ATU 123* and *ATU 333* that most likely evolved by blending together elements from both those types with local folktale motifs. The present study explores whether a phylomemetic approach can be similarly productive in resolving the typological questions surrounding another popular and much-debated international folktale: Cinderella.

2 The Cinderella Cycle

Cinderella is one of the most widespread and extensively studied stories in the international folktale record, with over 300 versions documented from around the world. Among them the earliest versions are the Egyptian version *Rhodopis* dating back to 1000 BCE and the Chinese version *The Story of Shen Hsien* reported by Alan Dundes (1988, 75) and dated back to the nineth century CE. Rather than comprising a single international type, these tales are generally thought to constitute a "cycle"

of inter-related tale types, the precise definition and classification of which varies according to the researcher. The first major attempt to catalogue Cinderella tales was carried out by Marian Roalfe Cox in the late nineteenth century (Cox 1893). Cox identified three main types of Cinderella, *Catskin – Cap o' Rushes*, *Cinderella*, and *Hero Tales*, along with a fourth group of miscellaneous *Indeterminate tales*. In a later study that incorporated a wider range of comparative material, Anna Birgitta Rooth developed a more comprehensive taxonomy that identified five main types, labelled A, B, AB, BI, and C (Rooth 1951). Rooth's classification intersects and overlaps with four of the Cinderella tale types listed in the Aarne-Thompson-Uther Index of International Tale Types (Uther, 2004) (see table 1). We will focus mainly on Rooth's typology, as it has the strongest empirical foundations and was explicitly developed in relation to historic-geographic hypotheses.

Tab. 1: The different Cinderella tale-types as categorised by different scholars

Type by Cox	ATU Type	Type by Rooth
Catskin		
Cap 'o Rushes	510 B	BI
Cinderella	510A	B
		AB
Indeterminate Tales	511	A
Hero Tales		C

We begin with a brief summary of Rooth's types, and their proposed relationships to one another.

Type A concerns an orphaned girl (or, in some Asian traditions, a brother and sister) who is persecuted by her stepmother and stepsisters. She is aided by a magical talking animal, who provides her with food or helps her complete an apparently impossible task. The animal is discovered by the protagonist's stepsisters, and killed. The heroine buries the animal's bones, from which grows a tree bearing riches or fruits. In some European versions, the girl meets a passing prince who stops to eat from the tree, and they get married. Type A is classified as 511 in the ATU Index. Cox did not include a named category equivalent to Type A, but listed a number of these stories as 'Indeterminate Tales' in her collection.

Type B is the "classic" Cinderella tale, where a mistreated heroine is set an impossible task by her stepmother and stepsisters to prevent her from attending a ball. A magical helper appears to undertake the task and gives the heroine magical

clothes. She goes to the ball where she meets the prince, but has to hurry home before the magic wears off and loses a slipper. The prince then searches the land to find the woman whose feet fit the shoe, and is eventually reunited with the heroine. The stepmothers and stepsisters are punished for their wickedness. Type B tales are classified as 510A in the ATU Index and as 'Cinderella' in Cox's typology. However, neither of those typologies separate Type B from Type AB (below).

Type AB comprises stories that share motifs with both Types A and B. The first act follows a similar plot as Type A, where a stepmother leaves the heroine to starve or who assigns her an impossible task. The heroine is helped by a magical animal that is discovered and killed. The hero buries the bones from which a tree grows and fruits magical clothes. The second act then follows Type B. The heroine wears magical clothes to a ball, where she disguises herself and meets a prince. She loses her shoe, which the prince uses to track her down and marry her. The stepmother is exposed and punished. In the ATU Index Type AB tales are split across 510A and 511, while Cox includes Type AB in her category of 'Cinderella' tales (with Type B).

Type C features a male hero, who is ill-treated by his stepmother and left to starve. He is helped by a magical animal, with whom he escapes through forests of metal before defeating monsters. His animal companion eventually perishes, giving the hero parts of his body, which he later uses to overcome a difficult task and marry the princess. This type corresponds to *ATU 511* and Cox's 'Hero Tales'.

Type BI, which includes Cox's tales of *Catskin and Cap o' Rushes*, is characterized by not being a stepmother story. The main motif of this type is the unnatural father, who condemns his daughter to death when she refuses to marry him, or fails to state her love for him. She escapes by disguising herself in an animal skin, and eventually finds employment in another palace as a servant. She attends three balls and meets the prince. Not realising her lowly status, he falls in love with the heroine and gifts her a special object, which she later uses to prove her identity and marry the prince, thus restoring her to royalty. Type BI corresponds to *ATU 510B*.

Rooth presents a rich and detailed analysis of the geographical distributions of each type and their associated motifs. Through a meticulous comparison of various "tradition areas", she builds a complex portrait of the evolution of the "Cinderella Cycle" that locates multiple centres of origin and directions of diffusion for the different types. For example, she proposes that Type A spread westward from the regions of India and Indonesia, eventually mutating into a specifically European subtype (AII) in which the sibling heroes are replaced by a lone daughter, who eventually marries a prince (an ending that is absent in the Asian sub-type AI). Similarly, Type C is seen by Rooth as an offshoot of Type A, in which a male hero is substituted for the persecuted daughter. She claimed that this innovation likely took place in the Middle East and then spread towards northern Europe, with variants now scattered across Scandinavia, Ireland, and in the Balkans. Whereas Types A and C

spread predominantly from east to west, Rooth suggests that Type B travelled in the opposite direction. Based on the density of variants of Type B, she argues that this type most likely originated in Europe, and probably migrated to Asia fairly recently. Type AB, on the other hand, presents a more puzzling case. As mentioned above, tales belonging to this type exhibit a miscellany of traits associated with Type A and Type B, and have been documented over a wide geographic area, from Ireland to Indonesia. Rooth considers two possible explanations for the origin of this type. The first sees Type AB as a hybrid form, created by blending together parts of the two other tale types. The second possibility is that AB is a transitional form, or evolutionary "missing link" between A and B that is younger than the former, but older than the latter. Rooth concluded that the second explanation is the more likely one, due to the paucity of Type B tales in Asia. In her view, the geographical distributions of the three types were more consistent with the theory that AB emerged from A, and then eventually evolved into B in European traditions. However, thanks to the work of the Chinese folklorist Nai-tung Ting (1974), it has since been discovered that Type B tales are a lot more common in Asia than Rooth realised, suggesting that this inference may be wrong, and that the hybridisation hypothesis cannot be discounted. Last of all, Rooth considered BI, which comprises the tales *Catskin* and *Cap o' Rushes*. Unlike Cox (1893) and Aarne-Thompson-Uther, Rooth concluded that this type, which is the only one lacking the wicked stepmother, is only distantly related to the other types, and represents a separate tale tradition.

Here, we aim to offer a new perspective on the evolution of the Cinderella Cycle by applying quantitative techniques from evolutionary biology to analyse cross-cultural variation in these tales. First, we test whether currently recognised types and sub-types of Cinderella tales represent phylogenetically distinct and coherent traditions. Second, we evaluate Rooth's historic-geographic models relating to the origins and relationships among the types.

3 Data Preparation

We compiled a sample of 266 recorded versions of Cinderella tale types representing a wide range of languages, cultures and geographical areas, albeit with a strong Eurocentric bias (Figure 1). A full list of tales and sources is provided in the Supplementary Information.

To code variations in the plot and prepare a dataset for analysis we deployed Dundes' concepts of the "motifeme" and the "allomotif" (Dundes 1962). Drawing on the distinction made in linguistics between morphemes (abstract units of meaning) and allomorphs (the various specific sounds or signs that express those meanings)

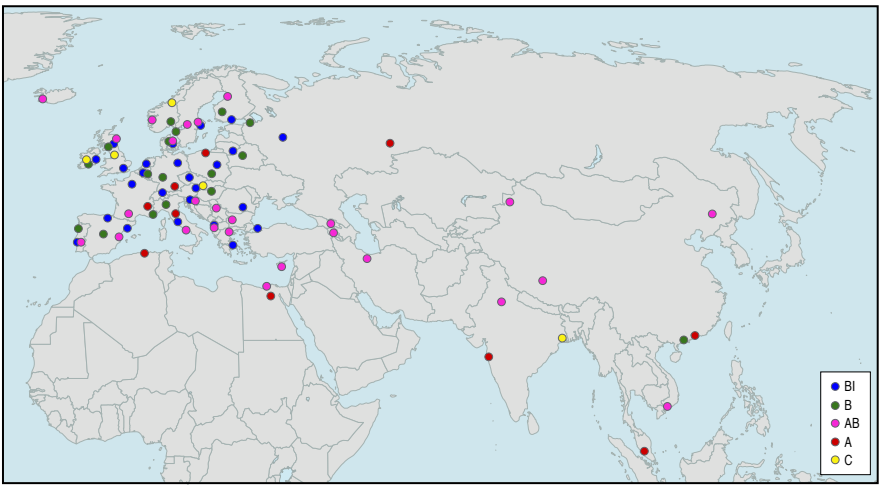

Fig. 1: Geographical distribution of Cinderella tale types among populations represented in the dataset

(Pike 1954), Dundes defines a motifeme as a structural component, or "function" (after Propp 1968), of a plot, while allomotifs represent the set of alternative motifs that express a given motifeme in a narrative corpus. For example, a key motifeme in Cinderella type tales is the "branding" of the heroine, whereby her identity is recognised by a special object. In different versions of the tale, this motifeme is realised through various allomotifs, including a lost shoe, a glove, a jewel and a ring, all of which fulfil the same role in advancing the plot. Dundes takes a synchronic, ahistorical approach to analysing allomotifs, employing them as a cipher to decode the underlying symbolic meanings and associations embedded in a corpus of folktales. Here instead, we adopt more of a diachronic perspective that utilises the substitution of allomotifs within motifemic slots as a means to track the transmission and mutation of stories over time. This strategy is analogous to the alignment of proteins or DNA sequences in genetics, which involves mapping the variant forms of genes (alleles) found across a series of structural positions (loci) in a genome.

We identified 74 motifemes in the tale sample based on Dundes' criteria – namely the *dramatis personae* and events or actions that advance the narrative (Dundes 1962). We then compiled a list of the motifs that occur in each motifemic slot – such as the branding of the hero through a lost shoe, glove, ring, etc. We term these "primary allomotifs". We then compiled a list of "secondary allomotifs" that describe variations in the primary allomotifs. For example, in the subset of tales that feature a lost shoe, the shoe is variously described as being made from glass, satin, silver, and so on (see Supporting Information for a full list).

We derived a total of 237 traits using this approach (see Supporting Information). The state of each trait was recorded for each tale in which it occurred and entered into a matrix. The absence of a trait was treated in one of two ways. "True absences" were defined as traits that could theoretically be present in a tale (e.g., because of the availability of multiple motifemic slots) but were not. This type of absence was recorded in the matrix and treated as a potentially informative state (i.e., as a trait that had either not evolved in the tale's lineage, or had been lost/replaced). "Bogus absences", on the other hand, were defined as traits that were not present because they contravened the narrative's logic, either because of the presence of another, mutually exclusive allomotif, or because of the prior absence of the corresponding motifeme or primary motif that it describes. For example, if the motifeme for the branding of the hero is absent, there can be no allomotifs for the slipper, glove, ring, etc. And if there is no slipper, then there can be no secondary allomotifs related to the material of which it is made (glass, silver, satin, etc.). These types of absences were recorded as "gaps" and treated as uninformative for the analyses. Accounting for the dependencies among traits, and downstream effects flowing from motifemes to primary and secondary allomotifs is an important feature of our coding approach. It explicitly recognises the syntagmatic structure of narratives rather than treating them as assemblages of independent traits, as previous phylogenetic analyses have done (e.g., Tehrani 2013, d'Huy 2015).

Prior to conducting any analyses of the dataset, we carried out a validation study that aimed to establish whether our coding method could reliably recover story transmission histories. This involved carrying out a series of transmission chain experiments designed to generate artificial tale lineages. The end points of the lineages were then coded using the approach described above and subjected to a phylogenetic analysis. The resulting phylogenies were able to reconstruct the known histories of the tales with an extremely high fidelity, and were more accurate than phylogenies reconstructed using a more standard coding approach. The validation study is presented in the Supplementary Information.

4 Methods

We employed three methods to investigate relationships among the tales included in our dataset: Bayesian phylogenetic inference, phylogenetic networks, and a model-based clustering method from population genetics (STRUCTURE). As we describe below, each method offers a distinctive approach to modelling evolutionary history, and by using them together we were able to explore a range of potential processes of diversification.

4.1 Bayesian phylogenetic inference

Bayesian phylogenetic inference (e.g., Huelsenbeck et al. 2001). aims to reconstruct relationships of common ancestry among a group of taxa – in this case, versions of Cinderella – by simulating their evolutionary histories as a branching process of descent with modification, whereby new lineages arise through the bifurcation of existing ones. It proceeds by calculating the likelihood of the data (i.e., the probability of obtaining the observed distribution of allomotifs among the tales) given an initial, randomly chosen, tree topology, a set of branch lengths (i.e., the evolutionary distances and amount of change separating ancestral tales and their descendants) and a model of trait evolution (i.e., the rates at which allomotifs mutate, and the variance of those rates across motifemes). The state of each of these parameters is then randomly modified (i.e., clades are re-sorted, branches get lengthened/shortened, variance in mutation rates is modified) and the likelihood of the data gets recalculated. This process is then repeated hundreds of thousands of times using a Markov Chain Monte Carlo (MCMC) chain algorithm. Moves that improve the likelihood of the data are always accepted, while those that do not are usually rejected (but have a small chance of being accepted to avoid the analysis getting trapped in local optima). Trees are sampled at regular intervals in the MCMC chain to compile a "posterior distribution of trees". Since the analysis usually favours moves that increase the likelihood of the data, trees with higher probabilities get sampled more often than ones with lower probabilities. The posterior distribution of trees can then be summarised by a consensus tree showing the relationships that are most frequently represented in the sample.

The analysis was carried out in the software programme MrBayes 3.2 (Ronquist et al. 2012) using the model settings for "standard" (multi-state, non-DNA) data, with the character coding set to "variable" and variance in rates of trait evolution estimated under a gamma distribution. Two analyses were carried out simultaneously, each using four MCMC chains with trees sampled every 1000 generations to avoid autocorrelation. After 5 million generations, a scatterplot of the log likelihood values of the tree samples indicated that the two runs had converged and the analysis was brought to a halt, with the first 25 % of each sample discarded as "burnin" (the exploratory phase of the MCMC search). A consensus tree of the remaining trees was then calculated on a majority-rules basis and the posterior probabilities for each clade (branch) estimated through the percentage of trees in which they were represented in the final sample.

4.2 Phylogenetic network analysis

The second approach we used, phylogenetic network analysis, captures conflicting relationships among a group of taxa. Although not as powerful as Bayesian phylogenetic inference for reconstructing relationships of common ancestry, this approach is useful when patterns of inheritance cannot be adequately accounted for by a strict branching model of evolution – for instance, when there is a significant transmission between, as well as within, lineages (Huson/Bryant, 2006). To construct a network for the Cinderella tales, we used the versatile NeighborNet algorithm, implemented in SplitsTree 4 (Huson/Bryant, 2015). NeighborNet first calculates pairwise distances between taxa, which represent the average number of mutations per trait that separate one taxon (e.g., a Cinderella tale) from another. The analysis then progressively partitions the taxa into a series of "splits", in which each taxon is paired with its nearest neighbour. When two pairs overlap (i.e., where the same taxon is represented twice), they are agglomerated to create two composite taxa. For example, if taxon A forms a pair with B and another pair with C, then each pair is agglomerated to form [A/B] and [A/C]. The distance of a composite taxon to all the remaining taxa is averaged from the two original taxa, and further splits are calculated. This process is repeated until a complete series of splits for the data have been obtained and no further agglomerations are possible. A key feature of the technique is that it allows taxa to be split in multiple, potentially conflicting, ways. These relationships are displayed in the form of a network, which shows groupings in the data (represented by parallel edges) and distances separating them (which are proportional to the lengths of the parallel edges). When the splits are highly consistent, as would be expected under a pre-dominantly phylogenetic (i.e., "vertical") model of evolution, the network will resemble a branching tree-like structure. Incompatible splits, on the other hand, produce box-like latticed structures, which indicate the operation of non-phylogenetic, or "reticulate" processes, such as borrowing ("recombination") and blending ("hybridisation") among lineages. The impact of these processes was quantified using the delta-score and Q-residual score (Holland et al. 2002; Gray et al. 2010). Both measures calculate conflicting signals by comparing path lengths among pairs of taxa on "quartets" (subsets of four taxa) selected from the network. Quartets are scored from 0 to 1 according to how resolved the splits between each pair of taxa are, with values closer to 0 being more tree-like and values closer to 1 more reticulate. The estimation of the delta score includes a normalisation constant, whereas Q-residuals had to be normalised by rescaling all between-taxa distances in the network so that they average 1. The NeighbourNet analysis and calculation of d-scores and Q-residuals were carried out in SplitsTree 4 (Huson and Bryant 2015)

4.3 A model-based clustering method from population genetics

The third technique we used was borrowed from population genetics. Whereas phylogenetics is concerned with mapping relationships between different species, population genetics usually focuses on the ancestry of localised groups or demes belonging to the same species. The reconstruction of relationships therefore involves modelling patterns of recombination rather than diversification ("speciation") and understanding how they are structured. In a genetic context, this entails determining whether the genotypic variation observed in a sample suggests that individuals show evidence of diverse ancestries or belong to a single, undifferentiated population (i.e., one consistent with an unbiased mating pattern, known as the "Hardy-Weinberg equilibrium", in which allele frequencies remain more-or-less constant from generation to generation). In the case of folktales, these techniques can test whether the variation exhibited by a set of tales indicate that their plots were built by combining and recombining allomotifs from a single, shared pool, or from historically distinct traditions. This makes it possible to identify the number of likely ancestral traditions, and quantify their relative contributions to individual versions in the tale sample. This is especially useful for investigating specific hypotheses about "admixture" (i.e., blending), such as the possibility that Cinderella tales classified as Type AB are descended from Types A and B.

We carried out our analyses using the software programme STRUCTURE 2.3.4 (Pritchard et al. 2000). STRUCTURE executes a model where the ancestry of each individual (in this case, each tale) is assigned to a fixed number of distinct source populations, K, so that the allele (or allomotif) frequencies within each ancestral population approximate expectations in the absence of mutation, migration or selection. This is achieved through a Bayesian MCMC procedure in which the composition of the putative ancestral populations is iteratively re-sorted thousands of times until the analysis achieves the best fit of the model to the data. We implemented a haploid, admixture model as we have one tale variant per sampled population and assume that each tale may derive from one or more of the K ancestral populations. We set up the model so that allomotif frequencies are assumed to be correlated among populations rather than varying independently, since the latter has a tendency to lump separate populations together if they share similar allomotif frequencies (Falush et al. 2003).

We ran a series of analyses under different values of K (i.e., the number of assumed ancestral populations of tales), running 10 replicates of each analysis to control for stochastic variations in the results. To determine appropriate run lengths for the MCMC chains we referred to STRUCTURE's diagnostic "alpha" sta-

tistic. Alpha values of <0.05 indicate that likelihood values for a given MCMC run have stabilised. On that basis, we ran the MCMC chain for each analysis for 150,000 iterations, with the first 50,000 iterations discarded as burn-in. We began with K=1 (where all tales would be assumed to belong to a single type) and progressively increased the value of K until the mean likelihood values plateaued, or showed high variance between runs, at which point it can be assumed there is no further population structure inherent in the data (Rosenberg et al. 2001). Last, we plotted the membership coefficients for each individual tale to elucidate the composition of the population clusters inferred from the analyses and explore how they map onto existing tale typologies and historic-geographic theories regarding the Cinderella Cycle.

5 Results of our analysis

5.1 Results of the Bayesian phylogenetic analysis

The results of the Bayesian analysis are summarised in Figure 2 below, which shows a majority-rules consensus tree calculated from the posterior distribution of trees. The tree is unrooted, so does not explicitly represent any chronological sequence or the direction of inheritance, meaning that the evolutionary histories of the traditions have to be interpreted from the structure and pattern of the relationships indicated by the tree. The tales are organised into phylogenetic groups known as "clans" (Wilkinson et al. 2007), which are equivalent to clades on a rooted tree. The clans shown in the consensus tree were represented in at least 50 % of the posterior distribution of trees. The comb-like appearance of the tree and shallow depth of most clans highlights a lack of agreement among the trees contained in the posterior distribution of trees, particularly with respect to larger phylogenetic groupings. This suggests the presence of significant conflicting signal in the data – e.g., because of hybridisation or horizontal transmission across lineages – and/or poor preservation of phylogenetic signatures, especially at deeper time depths.

Stories of Type B are labelled in green, Type A in red, Type AB in pink, Type C in yellow, and stories of Type BI, *Catskin* and *Cap o' Rushes*, are coloured respectively in dark blue and light blue. The spots on the interior branches are proportional in size to the posterior support for the corresponding clades.

The results provide little evidence to suggest that the currently recognised types of Cinderella tales comprise phylogenetically distinct traditions. Although tales

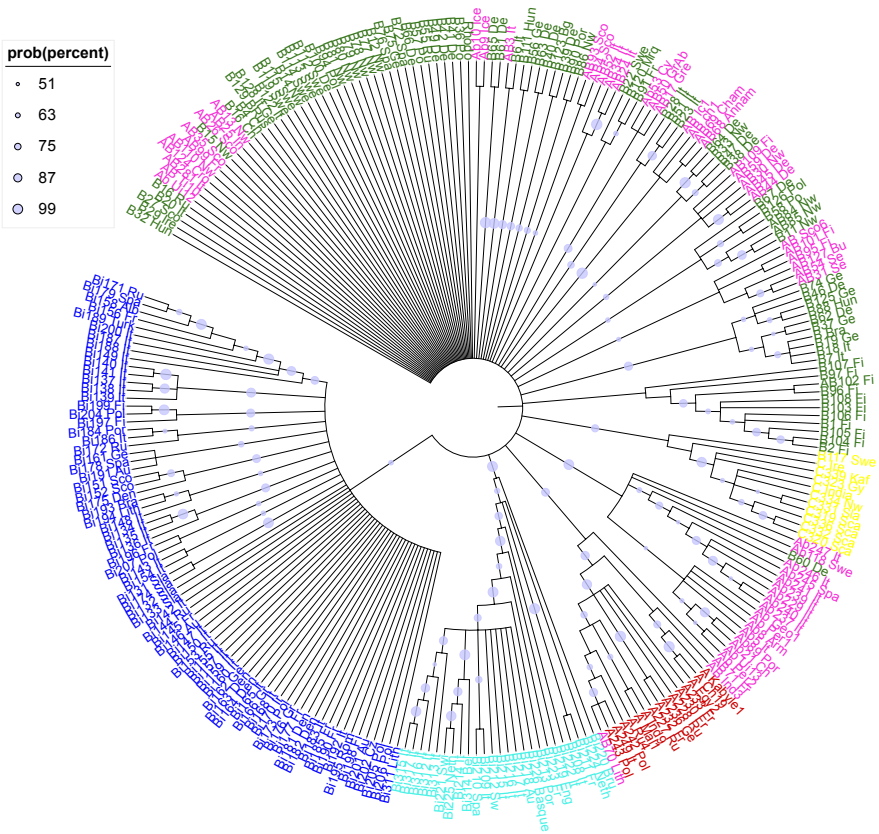

Fig. 2: Bayesian consensus tree of the five Cinderella types

often formed clans with members of the same type, the types tend to be highly fragmented. In general, they do not coalesce into larger groupings comprising all members of a single type while excluding other types. Only one of the types identified by Rooth fulfils this criterion, Type C, which formed a clan with a posterior probability of 100 %. Two other recognisable major clans consisted of tales belonging to Cox's types *Cap o' Rushes* and *Catskin*. The former group was strongly supported by the results, forming a clan with a posterior probability of 96 %. The latter was more modestly supported with a posterior probability of 59 %. However, there was no evidence to support Rooth's proposal that both groups belong to a common type (BI) that is distinct from other traditions of Cinderella. Tales belonging to Rooth's Type A, meanwhile, clustered together in a clan that had a reasonably high level of support (85 %). However, this clan also contained tales classified as Type AB

in Rooth's scheme, and a Type B tale. Tales classified by Rooth as Types B and AB were widely dispersed among various lineages, and did not form phylogenetically coherent groups.

The lack of phylogenetic structure, especially in deeper regions of the consensus tree, makes it difficult to evaluate historic-geographic hypotheses concerning the origins and relationships between types of Cinderella tale. Nevertheless, we must conclude that there is no evidence to support Rooth's hypothesis that Type C evolved from Type A. The hypothesis predicts that tales belonging to Type C should form a clan nested within the Type A, or to comprise an adjacent lineage to modern variants of Type A. Nor is there any evidence to support the hypothesis that Type AB is a transitional form between Types A and B. In that case, we would expect Type B to branch off from a Type AB clan, with the latter splitting from Type A. Instead, we observe some Type AB tales grouping with Type A, while others form clans with Type B tales. While this pattern is not consistent with the "transitional AB" hypothesis, it is potentially compatible with the alternative hypothesis that Type AB is a "hybrid" type that blended together motifemes and allomotifs from Types A and B, which could also explain the apparent lack of phylogenetic structure in the data. To investigate this possibility further, we turn now to the analyses that are better equipped to capture processes of "horizontal" transmission and blending than tree-based methods like Bayesian phylogenetic inference.

5.2 Results of the NeighborNet analysis

The network produced by the NeighborNet analysis (Figure 3) exhibits both tree-like and box-like patterns, with the latter being especially pronounced in the deeper structures of the network. The delta score and Q-residual of the network were 0.36 and 0.06 respectively, indicating the presence of a phylogenetic signal but with a significant degree of reticulation (e.g., Gray et al 2010; Tehrani 2013).

Stories of Type B are labelled in green, Type A in red, Type AB in pink, Type C in yellow, and stories of Type BI, *Catskin* and *Cap o' Rushes*, are coloured respectively in dark blue and light blue.

Relationships among tales are visibly more structured in the network than in the Bayesian consensus tree, with several clusters corresponding to recognised types of Cinderella. They include Cox's *Cap O'Rushes* and *Catskin*, which together form a larger grouping that equates to Rooth's Type BI (which was not present in the Bayesian tree). Rooth's Type A and Type C are also represented in the network and form adjacent clusters, which is consistent with her hypothesis that the latter

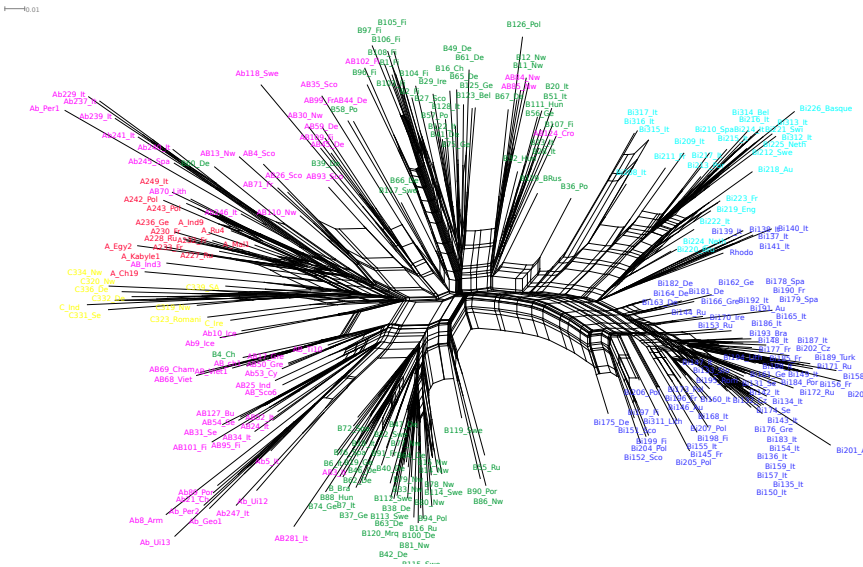

Fig. 3: NeighbourNet graph of the five types of Cinderella

developed as an offshoot of the former. Tales belonging to Type B, meanwhile, do not comprise a single group but are split into two main clusters, while Type AB tales are widely dispersed and exhibit complex and often conflicting affiliations with other tale clusters. Further exploration of these relationships reveals that the overlapping associations between Type AB with Type B on the one hand and Types A and C on the other have the effect of driving a wedge between Type B tales. When Type AB tales are removed from the analysis, Type B emerges as a clearly distinct cluster, while the relationship between Types A and C is brought into sharper focus. Overall, these results support Rooth's typology of Cinderella tales, while suggesting that Type AB is a product of hybridisation between the two distinct traditions, one comprising Type B tales, the other associated with the closely related Types A and C.

5.3 Results of the STRUCTURE analyses

The results of the simulations in STRUCTURE showed a steep increase in likelihood values from K=1 to K=2, rising further with each step up in the value of K before peaking at K=4, after which they decline and show high variance across separate runs (Figure 5). These results suggest that there is a discernible population struc-

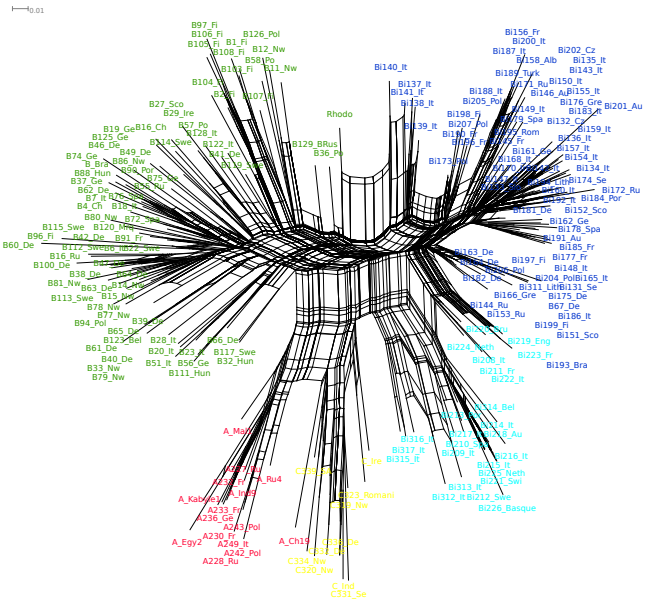

Fig. 4: NeighborNet graph with Type AB tales removed, with Types A (red), B (green), C (yellow), and BI (dark blue for Catskin, light blue for Cap O'Rushes) now forming distinct clusters

ture in allomotif frequencies, with up to four distinct ancestral populations contributing to the variation observed in our tale sample.

Figure 6 shows the estimated ancestry coefficients for tales belonging to each of Rooth's main types of Cinderella tale at K=2, K=3 and K=4. At K=2, there is a clear split between Rooth's Type BI and the other types A, B, AB and C. This result supports Rooth's hypothesis that Type BI tales belong to a separate tradition to the rest of the Cinderella Cycle. However, it is worth noting that this split is not absolute, with most tales exhibiting a degree of mixed ancestry. At K=3, the Type BI tradition is retained intact, while the tradition associated with Types A, B, AB and C splits into two. One of these is associated most strongly with Type B Tales, while the other tradition is dominant in Types A and C. This supports Rooth's contention that Types A and C are more closely related to each other than Type B. Type AB tales, meanwhile, present as a mixture of the two traditions. This pattern is consistent with the predictions of the "hybridisation" hypothesis for Type AB. It is not compatible with the alternative suggestion by Rooth that Type AB is a transitional form between A and B. If that were the case, the bar plots would be expected to separate Types A and C from Types AB and B, and for levels of admixture to be more even across all four types, rather than being so concentrated in

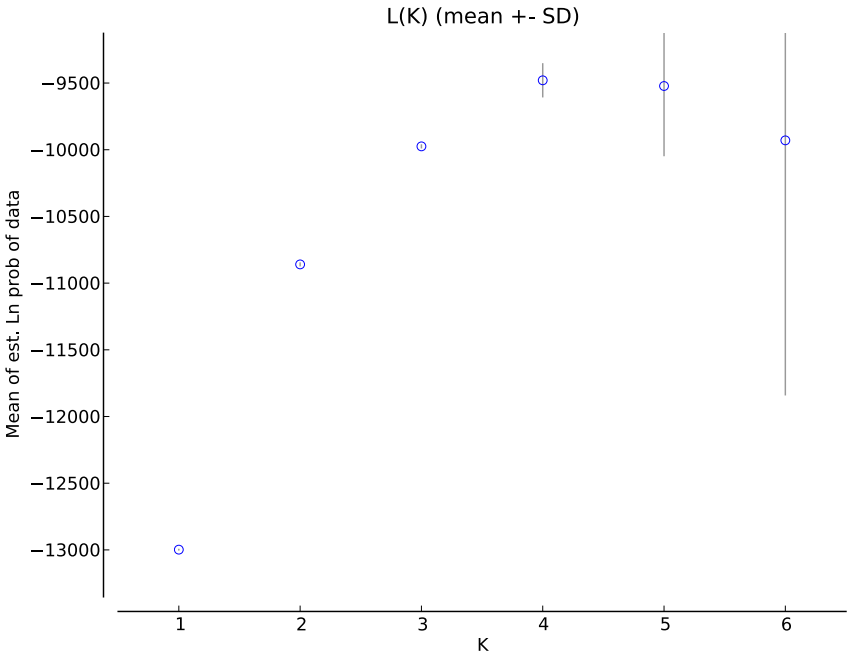

Fig. 5: Results of the STRUCTURE analyses, showing the mean likelihood of the data returned by different values of K (inferred ancestral populations)

one. At K=4, the Type BI tradition inferred from the K=2 and K=3 plots splits into two streams that reflect the distinction between *Cap o' Rushes* and *Catskin* tales. This new tradition also appears to have influenced a number of tales belonging to Type B, stirring a further ingredient into the cocktail of admixture uncovered by STRUCTURE.

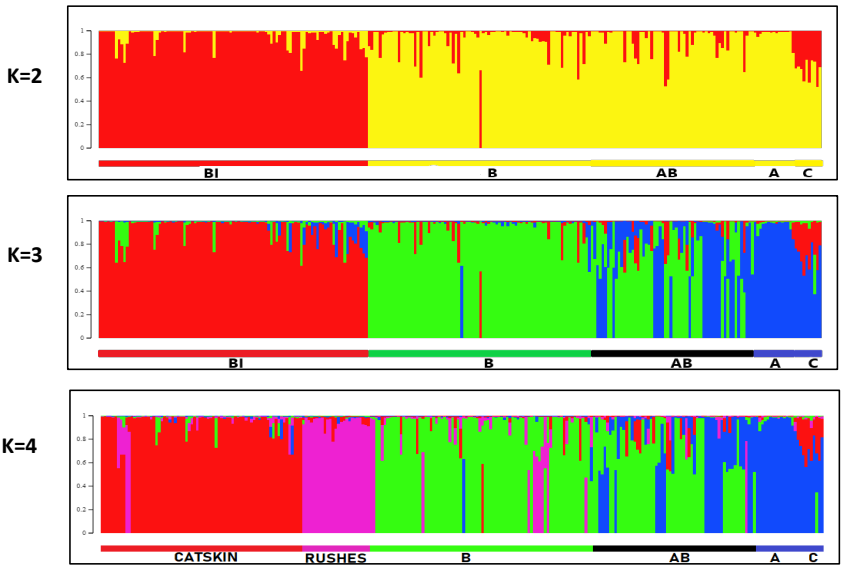

Fig. 6: Estimated ancestry coefficients for tales belonging to each of Rooth's main types of Cinderella tale at K=2, K=3, and K=4

6 Discussion

This case study has investigated two sets of questions concerning worldwide patterns of variation in the so-called "Cinderella Cycle" (*ATU 510–511*). First, to what extent do currently recognised types of Cinderella tale represent distinct and coherent traditions? Second, if such types do exist, what are their origins and relationships to one another? To address these questions, we developed an interdisciplinary approach that combined the structural analysis of folktales with quantitative evolutionary methods: First, we mapped 237 sites of variation in the plots of 266 versions of Cinderella based on Dundes' concepts of the "motifeme" and "allomotif" (Dundes 1962). We then carried out a series of "phylomemetic" (Howe/Windram 2011) analyses in which the variation observed across these sites was modelled as the outcome of cumulative processes of mutation/innovation and inheritance over time.

Overall, our results suggest that it is possible to identify distinct types of Cinderella tale, but show varying levels of support for each type across the different analyses we employed. Only two types, Cox's *Catskin* and *Cap O' Rushes*, were consistent with the results of all the analyses. The NeighborNet and STRUCTURE analy-

ses (though not the Bayesian phylogenetic analysis) further suggest the existence of a larger group comprising both these traditions, corresponding to Rooth's Type BI and the *ATU 510B* Type. Evidence relating to other tale types is more complex and, in some cases, contradictory. For instance, all the analyses support a separation between Rooth's Types A and B. However, distinguishing them from other types is far less straightforward. In the Bayesian and NeighborNet analyses, Type A and Type B were divided into clusters that also included versions of Type AB. They only formed exclusive groups once Type AB was removed from the dataset in a subsequent NeighborNet analysis. At first sight, these results might appear to be in line with the classification suggested by the ATU Index, which splits tales belonging to Rooth's types A, B and AB into just two types, *ATU 510A* and *ATU 511* rather than three (Table 1). However, as further analyses in NeighborNet (Figure 4) and STRUCTURE (Figure 6) showed, it is not as simple as that. Tales belonging to Type AB do not break cleanly between Type A and Type B but instead all show strong evidence of mixed ancestry. Last of all, Type C was strongly supported by the results of the Bayesian analysis, with these tales comprising an exclusive clan with a posterior probability of 100 %. Type C tales also clustered together in the NeighborNet analyses, especially following the removal of the conflicting signal associated with Type AB. However, STRUCTURE did not identify a unique ancestral signature in Type C tales that separated them from Type A. It is possible this may be due to the way that STRUCTURE pools tales into "interbreeding" traditions, rather than splitting them into progressively smaller and more exclusive lineages of descent, thereby differentiating groups at a lower level of resolution compared to phylogenetic trees and networks.

All of the major clusters returned by the analyses comprised a geographically and linguistically heterogeneous set of tales, which is consistent with the kind of long-range diffusion of tale types envisaged by historic-geographic theories. However, the lack of phylogenetic structure within each type, together with the uneven sampling of tales from different regions, prevents us from being able to locate their specific regional origins and pathways of diffusion. Nevertheless, it is still possible to draw a number of conclusions that are relevant to historic-geographic theories about the spread of Cinderella tales. First, Rooth's hypothesis that Type BI (*Catskin* and *Cap O' Rushes*) diverged from the other types at an early point in the evolution of these traditions is supported by both analyses. This is seen in the NeighborNet networks (Figures 3 and 4), which all show a clear split between BI and the other tales. In the STRUCTURE analyses the steepest increase in likelihood values occurred between K=1 (which assumes no population structure) and K=2 (Figure 5), with the latter model producing a sharp contrast in the estimated ancestry coefficients of Type BI tales versus Types A, B, C and AB (Figure 6). The STRUCTURE analyses further suggest that the two traditions that make up Type

BI, *Catskin* and *Cap O'Rushes*, likely emerged relatively recently and only appear at K=4, by which point Types A and B had already diverged from one another (K=3). The NeighborNet and STRUCTURE results are also consistent with another of Rooth's hypotheses, which proposes that Type C evolved from the same tradition as Type A. Both sets of tales cluster together in the NeighborNet networks and exhibit highly similar ancestry coefficients in the STRUCTURE plots at K=3 and K=4, which separate these tales from Types BI and B. The split between A and C appears to have occurred more recently than the origin of other Cinderella types, including the emergence of *Catskin* and *Cap O' Rushes* as separate traditions within Type BI. It is interesting to note that Type C, *Catskin* and *Cap O' Rushes* all formed distinct clans in the Bayesian consensus tree, suggesting a stronger phylogenetic signal in these comparatively young tales than in older types, where signatures of descent have been gradually eroded by time and the cumulative effects of borrowing and blending across lineages.

Our analyses also shed important light on the origins of Rooth's problematic Type AB. Rooth suggested two possibilities regarding these tales: either they represent a "transitional type" between A and B that has been preserved in some storytelling traditions, or a "hybrid" that blended together parts of the other two types. Our findings strongly favour the latter hypothesis over the former. Neither the NeighborNet or STRUCTURE results provide any evidence to suggest that Type B emerged from Type AB, or that Type A is ancestral to either of those types. In contrast, the NeighborNet analysis suggests that Type AB is a major source of conflicting signal in the data due to the affinities of these tales with both Type A and Type B. The STRUCTURE results, meanwhile, suggest that Types A, B and C all descend from a common tradition that split from Type BI, and later subdivided into two main streams, one leading to Types A and C, the other to Type B. The ancestry coefficients of Type AB tales indicate that these tales evolved by blending together these two lineages. A closer inspection of the NeighborNet and STRUCTURE results implies that this recombination of elements from A and B is likely to have occurred more than once, perhaps multiple times in various regions where Types B and A/C have come into contact or co-exist. This can be seen in the NeighborNet network, where Type AB tales do not form a single hybrid group that sits between a cluster for Type A and one for Type B, as would be expected if they had a single origin, but are instead distributed across multiple regions of the network, overlapping and intersecting with numerous subgroups of A and B with highly disruptive effects on the integrity of those types (Figure 4). Similarly, the ancestry coefficients for Type AB in the STRUCTURE plots show a lack of consistency, with the proportional influences of the two source traditions varying considerably across individual versions. Many of these tales also appear to draw on Type BI – a phenomenon noted by Rooth in relation to Scandinavian traditions especially (Rooth, 1951:224). Ultimately, these

findings suggest that Type AB should be considered to be a miscellaneous category of stories constructed from common source materials, rather than a coherent and distinct tradition in its own right.

While Type AB tales represent an extreme case, it is clear that recombination and admixture have played important roles in the development of all the Cinderella tale types. This presented significant challenges to the Bayesian phylogenetic analysis in particular. The consensus tree (Figure 2) returned by that analysis is lacking in structure and consists of small, fragmented clans. The tree offers no clues about deeper relationships of common ancestry at the level of different types. Far more informative results were obtained from the NeighborNet and STRUCTURE analyses, both of which are better able to capture processes of borrowing and blending that appear to be characteristic of Cinderella tale types. Indeed, it would appear that overall, these traditions are less sharply defined and phylogenetically coherent than other folktales that have been studied using phylomemetic methods. As mentioned previously, Tehrani was able to recover robust and distinct lineages of transmission for *Little Red Riding Hood* and *The Wolf and the Kids* (Tehrani 2013; Tehrani et al. 2016), while d'Huy found that cross-cultural variation in *Polyphemus* (d'Huy 2015) and *Pygmalion* (d'Huy 2013) can be captured by hierarchical, branching models of descent with modification. One other study by Ross et al. (2013) reported high levels of horizontal transmission among traditions of *ATU 480 The Kind and Unkind Girl*, but since their analyses focused on population-level patterns of diversity rather than relationships between specific versions or sub-types of the tale, their results are not directly comparable to the ones presented here.

What accounts for the relative fluidity of Cinderella traditions? One possibility is that our syntagmatic coding approach, which sought to account for structural dependencies among features of the stories (between motifemes and allomotifs, and between primary and secondary allomotifs), somehow obscured the phylogenetic signal in the data. However, this seems unlikely given the results of our validation experiment (see Supporting Information), which suggest that if anything our approach is *more* likely to recover evidence of common ancestry. In our view, a more plausible explanation is that there may be potent latent opportunities for borrowing and blending among this particular group of tales. Structurally, all the Cinderella types are all quite similar, with most of the observed variation among the stories occurring at the level of primary and secondary allomotifs, rather than motifemes (the latter accounting for only 74 out of the 237 traits used in the analyses). Given the large overlaps in the geographic ranges of the types (Figure 1), it is easy to see how storytellers might substitute allomotifs from one tale type for another when they share a common set of motifemes. When it comes to Cinderella motifs, it would appear that the same shoe can fit many feet.

7 References

Aarne, Antti/Thompson, Stith: The Types of the Folktale. A Classification and Bibliography. Helsinki 1961.

Cox, Marian R.: Cinderella. Three hundred and forty-five variants. London 1893.

d'Huy, Julien: A phylogenetic approach of mythology and its archaeological consequences. Rock Art Research. In: Australian Rock Art Research Association 30 (2013) 115–118. ffhalshs-00932214.

d'Huy, Julien: Polyphemus, a Palaeolithic tale? In: The Retrospective Methods Network Newsletter 9 (Winter 2014–2015) 43–64.

Dawkins, Richard: The selfish gene. New York 1976.

Dundes, Alan: From etic to emic units in the structural study of folktales. In: The Journal of American Folklore 75,296 (1962) 95–105.

Dundes, Alan: Cinderella, a casebook. Vol. 3. Wisconsin 1988.

Dundes, Alan: The Motif-Index and the Tale Type Index. A Critique. In: Journal of Folklore Research 34 (1997) 195–202.

Falush, Daniel/Stephens, Matthew/Pritchard, Jonathan K.: Inference of population structure using multilocus genotype data. Linked loci and correlated allele frequencies. In: Genetics 164 (2003) 1567–1587.

Goldberg, Christine: The Historic-Geographic Method: Past and Future. In: Journal of Folklore Research 21 (1984) 1–18.

Gray, Russell D./Bryant, David/Greenhill, Simon J.: On the shape and fabric of human history. In: Philosophical Transactions of the Royal Society 365 (2010) 3923–3933.

Holland, Barbara R./Huber, Katharina T./Dress, Andreas, Moulton, Vincent: δ Plots: A Tool for Analyzing Phylogenetic Distance Data. In: Molecular Biology and Evolution 19 (2002) 2051–2059.

Howe, Christopher/Windram, Heather: Phylomemetics-Evolutionary Analysis beyond the Gene. In: PLoS Biol 9 (2011) e1001069.

Huelsenbeck, John P./Ronquist, Frederik/Nielsen, Rasmus/Bollback, Jonathan P.: Bayesian inference of phylogeny and its impact on evolutionary biology. In: Science 294 (2001) 2310–2314.

Huson, Daniel H./Bryant, David: Application of Phylogenetic Networks in Evolutionary Studies. In: Molecular Biology and Evolution 23 (2006) 254–267.

Huson, Daniel H./Bryant, David: User manual for SplitsTree4 V4.4. 2015.

Jacobs, Melville: A Look Ahead in Oral Literature Research. In: The Journal of American Folklore 79 (1966) 413–427.

Pike, Kenneth L.: Language in Relation to a Unified Theory of the Structure of Human Behaviour. Part I. Glendale 1954.

Pritchard, Jonathan K./Stephens, Matthew/Donnelly, Peter: Inference of population structure using multilocus genotype data. In: Genetics 155 (2000) 945–959.

Propp, Vladimir: Morphology of the Folktale. Second Edition, Revised and Edited with Preface by Louis A. Wagner, Introduction by Alan Dundes. Austin 1968.

Ronquist, Frederik et al.: MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. In: Systematic Biology 61,3 (2012) 539–542.

Rooth, Marian R.: The Cinderella Cycle. Lund 1951.

Rosenberg, Noah A. et al.: Genetic Structure of Human Populations. In: Science 298, 5602 (2002) 2381–2385.

Ross, Robert M./Greenhill, Simon J./Atkinson, Quentin D.: Population structure and cultural geography of a folktale in Europe. In: Proceedings of the Royal Society B: Biological Sciences 280 (2013) 20123065.

Sydow, Carl W.: Selected Papers on Folklore. Copenhagen 1948.
Tehrani, Jamshid J.: The phylogeny of little red riding hood. In: PLoS ONE 8,11 (2013).
Tehrani, Jamshid J./Nguyen, Quan/Roos, Teemu: Oral Fairy Tale or Literary Fake? Investigating the Origins of Little Red Riding Hood Using Phylogenetic Network Analysis. In: Digital Scholarship in the Humanities 31,3 (2016) 611–636.
Tehrani, Jamshid J./d'Huy, Julien: Phylogenetics meets folklore. Bioinformatics approaches to the study of international folktales. In: Maths Meets Myths. Quantitative Approaches to Ancient Narratives. Ed. Ralph Kenna/McCarron, Máirín/McCarron Pádraig. Heidelberg 2017, 91–114.
Thompson, Stith: The Folktale. New York 1951.
Ting, Nai-tung: Cinderella Cycle in China and Indo-China. Helsinki 1974.
Uther, Hansjörg: The Types of International Folktales. A Classification and Bibliography. Parts I–III. Helsinki 2004.
Wilkinson, Mark et al.: Of clades and clans: terms for phylogenetic relationships in unrooted trees. In: Trends in Ecology & Evolution 22,3 (2007) 114–115.