The Marlowe-Shakespeare Authorship Debate:

Approaching an Old Problem with New Methods

Albert C.-C. Yang, C.-K. Peng, Ary L. Goldberger

Margret and H.A Rey Institute for Nonlinear Dynamics in Medicine, Beth Israel

Deaconess Medical Center/Harvard Medical School, Boston, Massachusetts 02215, USA

Correspondence

Ary L. Goldberger, MD

Director, Margret and H.A Rey Institute for Nonlinear Dynamics in Medicine

Beth Israel Deaconess Medical Center (GZ-435)

330 Brookline Avenue, Boston, MA 02215

Tele: 617-667-4267; Fax: 617-667-4012

e-mail: agoldber@bidmc.harvard.edu

"We must look for consistency. Where there is want of it we must suspect deception."

Sherlock Holmes in *The Problem of Thor Bridge*

I. Introduction: The Shakespeare Authorship Problem

William Shakespeare, poet and dramatist, has become the pre-eminent symbol of literary genius. Nevertheless, the author's "real" identity has been an inexhaustible source of controversy and mystery for over 300 years, with more than eighty Elizabethans having been proposed since the eighteenth century as the "true Shakespeare." Among the many candidates, Shakespeare's contemporary, Christopher Marlowe, has attracted much attention since 1895 when William Gleason Zeigler published a detective story entitled *It Was Marlowe: A Story of the Secret of Three Centuries*[1].

The facts surrounding the life and death of the men called Shakespeare and Marlowe are murky at best. Both men had births recorded in 1564. Before Shakespeare's name became widely known, Marlowe had already produced several major works in various genres, including *Tamburlaine the Great* and *Dr. Faustus*. According to conventional accounts, Marlowe's career tragically ended on 30 May, 1593 when he was apparently murdered in a dispute. The mysterious circumstances surrounding Marlowe's death and the coincidental appearance at around the same time of a young actor and playwright named William Shakespeare have helped fuel speculation that Marlowe, in fact, may not have

been killed. Indeed, in 1955, Calvin Hoffman published a book entitled *The Murder of the Man who was Shakespeare*[2] in which he elaborated the intriguing hypothesis that Marlowe had not died as claimed in 1593, but instead escaped to a secret refuge in Italy where he spent the rest of his life writing the body of plays generally attributed to Shakespeare.

Another, less controversial possibility raised by others, is that Marlowe, while not the author of the entire Shakespearean canon, may have heavily influenced or even written or contributed to some of Shakespeare's early plays, including the *Henry VI* trilogy*, Richard II*, and *Richard III*[3-5]. The lack of proper documentation of most plays in the late 16th and early 17th centuries has compounded debates about the authorship of these and other plays. *Edward III*, for example, has been controversially attributed both to early Shakespeare and to Marlowe[6-9], as well as to others[10].

Given the disputed historical records and the lack of any solid traceable "evidence" other than the existing versions of the dramas, how does one proceed to analyze scientifically and objectively the problems of Shakespearean play authorship? In keeping with the Holmesian admonition in the opening quote, we believe that a fruitful approach to text analysis should rigorously and quantitatively compare the *consistency of styles* between plays by two putative authors. In this essay, we describe a new approach to assessing stylistic consistency between Shakespeare and Marlowe. The body of this paper, therefore, is devoted to a brief overview of scientific approaches to authorship identification, followed by a description of our new approach based on information categorization – *word rank*

order and frequency analysis—and, finally, the application of this new technique to the Marlowe-Shakespeare controversy.

II. Authorship Attribution: An Overview

Strings of words encode the dynamic signature of an author's language, just as genetic sequences (DNA) encode a species' features. Literary experts are often able to reliably discriminate between the styles of contemporary authors who may sound quite similar and even indistinguishable to general audiences. However, this type of expert diagnosis, although powerful, is subjective and often non-quantifiable. Recently, therefore, considerable effort has been devoted to the development and application of fully objective techniques to compare the stylistic features of literary texts.

The first attempt at a quantitative study of authors' styles can be traced back to 1887 when Thomas Mendenhall studied the statistics of word lengths[11] (number of letters in a word). He demonstrated that Francis Bacon was not likely to have written Shakespeare based on an analysis of such statistics. Similarly, Yule[12] (1938) and Morton[13] (1965) investigated the statistics of sentence lengths. However, both approaches to the Shakespeare authorship problem were inconclusive.

In 1964, Mosteller and Wallace[14] published an influential work on authorship attribution which introduced the concept of comparing the frequencies of *function words*, defined as non-contextual words such as "and", "the", or "to". In contrast to contextual words, function words are likely to appear commonly in all texts. Mosteller and Wallace investigated the usage of the function words using a specific statistical method (Bayes' theorem) and concluded that James Madison was the author of all twelve disputed essays in

the *Federalist Papers* published in the late 18th century in support of the United States proposed new Constitution[15]. In the late 20th century, additional linguistic features and statistical techniques were investigated to address the vexing problem of disputed authorship claims[16-24].

Of note, most of the contemporary statistical techniques developed for forensic text analysis focus on specific linguistic features, especially word[11] or sentence lengths[12,13], grammar[22], function word frequencies[14,16,23,24], and vocabulary richness[18]. However, these techniques are limited because they require specialized knowledge to identify the proper set of words for analysis and because they are not applicable to the wide range of written human languages. For example, Chinese and the related family of Asian languages have no word length, function words are language-specific, and many words have substantially permuted their meaning over the past few centuries.

On a more fundamental level, the difficulties of authorship attribution and the limitations of current techniques originate from the abstract nature of the *information* contained in literary texts. Each author is a unique source of information. *Therefore, the challenge of authorship attribution can be regarded as a special case of the problem of information categorization*. This challenge, in turn, should be approachable by studying generic features of information-carrying sequences. In the next section, we will describe a simple, yet generic statistical measure for information categorization that we have developed for this purpose. We then demonstrate its versatile applicability to forensic text

analysis of English as well as non-English literary works, focusing on its specific application to the Shakespeare-Marlowe debate.

III. What is Information?

Everyday life is full of *information*, ranging from items in newspapers or televised reports, to highway signs, to works of music and literature. A common underlying feature of the many different types of information-carrying entities is that they can be coded and broadcast by a sequence of symbols. Symbolic sequences as information carriers are, indeed, universal in nature. Some of these sequences are uniquely human creations, such as language, both spoken and written, and music. Other information sequences are generated by natural processes, such as genetic (DNA) codes and the electrical transmission signals of our nervous systems. The central problem in the analysis of these complex sequences is how to effectively categorize their origins based on their information content. For example, musical compositions by different composers usually display differentiable styles that can be recognized by experienced listeners. Unfortunately, this type of categorization that is processed by complex cognitive processes is not yet quantifiable.

In order to develop an effective algorithm to solve this challenge, it is useful to begin with perhaps the most fundamental, common feature of information-carrying sequences, namely, that the information is encoded by a finite number of *repetitive elements*. This property is immediately apparent for all written and spoken human languages, in which the repetitive elements are words. (Of course, when there is new information not able to be described, a new "word" must be created.) Furthermore, other forms of "words" or their equivalents can also encode different types of information. For example, the genetic sequences of the millions of species on the earth are simply based on the arrangements of four basic molecules, called nucleotides, abbreviated as *A*, *G*, *C*, and *T*.

This common feature supports the intuition that the composition of "words" is closely related to the information carried by the sequence itself. Therefore, information categorization should be made possible by comparing word compositions. To proceed further with developing our algorithm, we need to demonstrate two generic properties when comparing symbolic sequences.

The first property is that each symbolic sequence has a unique preference for a specific set of words. Closely related sequences are more likely to choose words in a similar way. This principle is obvious in the case of language. For example, although it is possible to recognize all of the words in a given language, not every word will appear in a given text. Instead, each author has his/her own vocabulary "database," related to education, culture, and life experiences. Further, when comparing the writings of two different authors, there will be *shared words* used by both, as well as *unique words* used specifically by one author but not the other. The number of shared words represents the degree of vocabulary overlap among different texts. An author may write on considerably different topics over the course of a career. In comparing one author's texts to those of others, however, the number of shared words is usually higher between texts by the same author than between texts by different authors, indicating the importance of word preferences.

The second property is that each symbolic sequence also has a unique preference for the frequency with which words are used. A word frequently used by one author may be used much less often by another author. This phenomenon has already been validated with

respect to *function words*. (i.e., non-contextual words) as described above. However, we have discovered that this property is not limited to the small set of *function words* as proposed by previous studies[23,25], but rather is a generic feature of all *shared words* when comparing two texts.

These two generic properties are closely related to the problem of information categorization. The information (meaning) conveyed by a word is related to how frequently it appears in a text. For example, commonly used words such as "the", "and", or "to" have the least information but are used most frequently. In contrast, rarely used words usually convey very specific information and are not likely to be used throughout the text. The hierarchy of word usage, therefore, represents how an author uses the information content of words. Since each author may learn words differently since childhood, the word preference should be a reasonable characteristic of an author's "style." Therefore, measuring stylistic similarity should be made possible by comparing the difference in frequency of shared words employed by different authors.

The two generic properties described above and their connections to the information categorization problem give us the basis to formalize the distance (or dissimilarity) between symbolic sequences. The technical details of this recipe (algorithm) have been previously published[26] and are also summarized in the Appendix. The basis of this approach, however, can be readily grasped through a graphic representation.

To visually illustrate our word rank order and frequency algorithm for measuring the distance between texts, we start with an example, selecting two of Shakespeare's plays,

Cymbeline (1609) and *The Winter's Tale* (1610), as well as one of John Fletcher's plays, *Bonduca* (1611). The first step is to construct a word frequency list (see Figure 1) by counting the occurrence of each word in a text, and ranking them by descending frequency (most frequent to least frequent). Next, we identify those words that are *shared* by the two texts. The rank order differences between two texts can then be visualized by plotting the rank number of each shared word in the first text against that of the second text. For example, the word "me" is the 20th most frequently used word in *The Winter's Tale*, but the 11th most frequently used word in the play *Bonduca*. Therefore, in Fig. 1b, the word "me" is placed at a location with horizontal coordinate of 20 and vertical coordinate of 11. If the rank-frequency statistics of two texts are identical, the shared words will fall along the diagonal line connecting the lower left and upper right parts of the graph. Figure 2 shows the comparison between the two Shakespeare plays. The words are tightly centered along the diagonal line, indicating the rank of each word is very similar in these two plays. In contrast, when the same comparison is made between the plays attributed to Shakespeare and the Fletcher play, respectively, the words are more widely scattered around the diagonal "line of identity."

As demonstrated by the above examples, the "distance" (or dissimilarity) between any two texts can be quantified by measuring the scatter of these points from the diagonal line in this rank comparison plot. Greater distance indicates less similarity and vice versa. Based on this information categorization approach, we have developed a quantitative index to measure the dissimilarity between symbolic sequences[26,27] (see Appendix for technical details). In the next section, we illustrate how to combine this simple distance metric based

on the rank and frequency of shared words with a *phylogenetic analysis technique* designed to reconstruct a classification "tree" of texts and uncover their authorship identities.

IV. Phylogenetic Analysis

The origin of phylogenetic analysis can be traced to Charles Darwin[28], who discovered the "pathway of species," termed evolution, in 1859. Darwin believed that the comparative study of the anatomical structures of living animals was crucial for reconstructing the "tree" of extinct and existing species. Less complex species would be located near the root and higher level animals would be located at the top of the tree. Today, this idea has been widely applied to phylogenetic studies based on the elucidation of the structure and function of the DNA molecule by Watson and Crick[29]. The discovery has made it possible to analyze the genetic codes letter by letter and objectively reconstruct the phylogenetic tree of different species[30-32].

The critical step in phylogenetic analysis is to estimate the similarity between pairs of species. One can then use these similarity data to arrange closely related species on nearby branches, while placing dissimilar pairs on more distant branches[31]. For example, based on the similarity of genetic codes, chimpanzees are more similar to humans than gorillas are, indicating the closer evolutionary connection between the two former species, with gorillas being more ancient ancestors. Therefore, we can place the gorilla at the root of the three-species phylogenetic tree, and then reconstruct the correct evolutionary pedigree of its primate relatives.

In a similar fashion, based on the simple similarity word rank order and frequency index described in the previous section, we can also construct a "tree" of literary texts to

uncover their identities. First, we need to determine the similarity between pairs of texts and group them onto a proper branch. If an unknown text is then placed on a known author's branch, the text is more likely by that author and not by the other candidates. Furthermore, by assigning the proper root, namely the very first text of a given author, we will be able to reconstruct the "evolutionary" tree of an author's writing chronology.

Figures 1 to 3 illustrate our automated computer algorithm for comparing literary texts which comprises the following three steps: 1) construct the word frequency list for each text and rank each word according to its descending frequency; 2) compare the similarity between each pair of word frequency lists by the rank order difference of shared words, and 3) use the phylogenetic algorithm to reconstruct the tree of the texts by grouping similar texts on nearby branches and dissimilar texts on distant branches. For the example given in Figure 3, we construct a mini-database containing five of Shakespeare's late plays, four of John Fletcher's sole-authorship plays, as well as two of their putative collaborations, *Henry VIII* and *The Two Noble Kinsmen*. The phylogeny based on the pairwise distance of these plays successfully recovers their widely held relationships; namely, the plays attributed to Shakespeare are arranged on different branches from those of Fletcher, with the Shakespeare-Fletcher collaborations falling between their respective branches. These computations can performed with great rapidity: processing the entire Shakespearean canon can be completed within 10 minutes on a laptop computer.

In the next two sections, we first present a further validation of our word rank order and frequency algorithm on various texts by known authors, and then briefly show its use in

the *Federalist Papers* dispute, as well as in the authorship debate surrounding the classical Chinese novel, *The Dream of the Red Chamber*. After validating the accuracy of the method and its applicability to very different languages, we turn our attention to the identities of Shakespeare and Marlowe, as well as to the controversial authorship of the anonymous play, *Edward III*.

V. Validation

1. Multiple Authorship Test for English Texts

For initial validation of our new inter-textual distance measurement, we first applied it to a database containing sixteen texts by eight well-known authors[18], including two works by each author (except for a single work from Emily Brontë and three from Sir Arthur Conan Doyle). The texts were obtained from the Oxford Text Archive and varied in length from *The Acts of the Apostles* with 24246 words, to the 116534 words that comprise *Wuthering Heights*. This dataset allowed us to examine the accuracy of our distance metric. We found that the method can unambiguously classify all of authors without error (Figure 4). In comparison, methods proposed by previous studies[17,18,20,23,33] for authorship identification misidentified at least one author.

2. The Federalist Papers

Next, we applied our method to authorship problems concerning the *Federalist Papers*[15,34], a series of 85 essays written on the proposed new U.S. Constitution and the nature of republican government. The Federalist Papers are believed to have been written between 1787-88 by Alexander Hamilton, James Madison, and John Jay[15]. Since 1788, the consensus has been that Alexander Hamilton was the sole author of fifty-one Papers, John Jay of five, James Madison of fourteen, and that Hamilton and Madison collaborated on another three. The authorship of the remaining twelve Papers (numbers 49-58, 62, 63) has been in dispute. Analysis using the work order rank frequency method supports Mosteller

and Wallace's as well as others' conclusion that Madison was the author of all twelve disputed Papers[25,35,36].

3. Chinese Literature: The Dream of the Red Chamber

We also applied the new distance measure to a very different language – Chinese. This and related ideographic languages use characters to represent "words." The very different characteristics of Chinese ideograms make it impossible to analyze them by conventional linguistic techniques based on Western language features. We first validated the method on a small database of Chinese literary works with known authorship. We then studied a controversy surrounding the authorship of the classic 18th century novel *The Dream of the Red Chamber*[37], one of the most influential texts in Chinese literature. However, the original manuscript was lost and only incomplete hand-written copies have survived. A century-old debate centers on the consistency of the first 80 chapters and last 40 chapters[38]. We found that the distance between the first 80 and the last 40 chapters is generally larger than the distance within both parts. This result is consistent with the scholarly consensus that the novel was composed by two primary authors[38].

VI. Re-examining the Shakespeare-Marlowe Controversy

Having validated our algorithm on three independent databases, we now consider the central issue regarding the Shakespeare-Marlowe authorship problem. To demonstrate how our method may provide some useful insights to this challenge, we constructed a database of 45 plays, including those currently in Shakespeare's canon as well as works by Marlowe. The Shakespeare texts are based on the First Folio[39,40] and the Globe edition[40,41]. The former collection, containing thirty-six plays (except *Pericles*), appeared in 1623, seven years after Shakespeare's death. The texts of First Folio plays are in old style spellings and punctuations. The later Globe edition published in 1866, contains thirty-seven plays (including *Pericles*) and several poems with contemporary spelling. In addition, we also collected seven of Marlowe's plays in both old and modern versions[39,42].

Stylistic Consistency Test Between Shakespeare and Marlowe

We first divided the plays into two groups of old and modern versions in order to maintain consistency of spelling between the Shakespeare and Marlowe texts. We then calculated the distance between each pair of plays using the word rank order and frequency method described above and detailed in the Appendix. The resulting phylogenetic tree of these texts based on the modern versions is shown in Figure 5. Of particular note is that all of the Marlowe texts are arranged on a very different branch from Shakespeare's plays, suggesting two different authors for the Shakespeare and Marlowe works. We do observe that the early historic plays by Shakespeare, including the *Henry VI* trilogy, are arranged closer to the Marlowe branch, consistent with his purported influence on these early

Shakespearean works[3-5]. We therefore conducted a further analysis comparing Shakespeare's early plays with Marlowe's texts using the old versions. The result is shown in Figure 6. The distance between the early Shakespeare and Marlowe texts is even closer than that in Figure 5. However, their works are still distinct without overlap. *These robust statistical findings indicate that the authorship of these early Shakespeare plays cannot be attributed to Marlowe, but at the same time, support the hypothesis that Marlowe did have an important influence on Shakespeare's works during this formative phase of his career.*

A Controversial Play: *The Raigne of King Edward III*

The play *Edward III* entered the stage anonymously on December 1, 1595[43]. It did not attract much attention until Edward Capell, a noted 18th century Shakespearean editor, included this play in his book *Perimedes the Blacksmith*[44]. He proposed that the play was "written by Shakespeare" based on the "integrity" of the play with Shakespeare's early writings. Since then, the authorship of the play has been a matter of dispute[6-8,45-47]. The play has recently been controversially restored to the Shakespearean canon[8]. Although the role of Shakespeare in authoring this historical drama is not established, several scholars have speculated that Shakespeare may have witnessed the battle of the Spanish Armada alluded to in the play[9]. Moreover, the play contains three direct quotes from Shakespeare's sonnets, indicating the author (or co-author) of the play either knew Shakespeare's works or was even Shakespeare himself[7,8]. However, Marlowe's contribution to the play has also been seriously considered[48-50].

We analyzed the Shakespeare and Marlowe texts in both old and modern versions (Figures 5 and 6). Both results show that *Edward III* is classified under the Marlowe, not the Shakespeare, branch. This finding strongly supports the contention that Marlowe did the majority of the writing on *Edward III*[48-50].

Next, we investigated the problem of dating the play using the phylogenetic approach. There are two widely-cited reference points for dating *Edward III*[51]: 1) the battle of the Spanish Armada occurred in 1588, and 2) the play publicly appeared in 1595. Therefore, the dating of the play likely lies between these dates. To construct the chronology of Marlowe's canon, we need to assign the root of the tree, that is, his very first play. Although the chronology of Marlowe's canon is also in great dispute, the consensus of scholars favors *Tamburlaine the Great* as the first play, dated around 1587[52]. We can, therefore, assign this work as the root of the tree and the resulting phylogeny is shown in Figure 7. *Edward III* falls between *Tamburlaine the Great, Part 2* (1588) and *Dr. Faustus* (1589-90), suggesting the date of the disputed former play around 1588-90.

The Chronology of Shakespeare's Canon

The chronology of Shakespeare's early plays is also uncertain[3,4], but a reasonable approximation of their order can be inferred from dates of publication, references in contemporary writings, historical events, and stylistic comparisons. His first plays are believed to be the three parts of the *Henry VI* series. Two parts of the trilogy *(2 and 3)*, were printed around 1594 under very different, lengthy titles. The first part of *Henry VI* did not appear until the First Folio was published. Wells and Taylor[53] proposed that *Henry VI,*

Part 2 was written around 1590, followed by *Henry VI, Part 3* in 1591 and *Henry VI, Part 1* in 1592. Therefore, we chose *Henry VI, Part 2* as the root of Shakespeare's canon and constructed the overall phylogenetic tree based on that assumption.

The resulting Shakespeare dramatic phylogeny (Figure 8) reasonably places the early plays (1590-94), middle plays (1595-1601), and late plays (1602-1613) from the bottom to the top. Furthermore, the tree clearly shows an evolutionary progression in Shakespeare's early plays including *Henry VI, Parts 1, 2, 3*, *Richard II*, *III*, *King John*, and *Titus Andronicus*. These early histories were written around 1590-94 and are considered to be heavily influenced by Marlowe[3-5]. In contrast, all of the subsequent plays, including the early comedies, lie on different but adjacent branches and display a remarkable internal consistency, suggesting a post-Marlovian phase in the author's stylistic evolution.

VII. Discussion

Who wrote "Shakespeare" is a question that has intrigued scholars and general readers for hundreds of years. Assuming Shakespeare and Marlowe were the same person, as proposed by Hoffman, Marlowe would have to have lived past 1593. However, our results using a validated statistical approach to establishing textual phylogeny show that Marlowe's texts are classified on a separate branch clearly distinct from the Shakespeare texts. The gap between these plays is too large to be accounted for by a change in style, even assuming Marlowe survival. The results, therefore, strongly support the view that Marlowe is not the author of the body of Shakespeare's plays.

For the disputed play, *Edward III*, however, we found a very different pattern of word usage from Shakespeare's major works, suggesting that the play was indeed written primarily by someone else. Of note, *Edward III* falls on a tree branch far from Shakespeare, and is grouped with Marlowe's works, indicating a stylistic similarity between the play and works generally attributed to the Canterbury dramatist. Our analysis is consistent with one recent scholar's suggestion that Marlowe did the majority of the writing on *Edward III*[48-50]. However, our findings do not exclude substantial editing by others or even partial contributions by Shakespeare, as suggested by some scholars[7-9].

Furthermore, the evolutionary pattern in the early plays before 1594 suggests that the young Shakespeare rapidly evolved his writing style toward its mature and peerless apogee. However, the narrower distance between Shakespeare's early plays and Marlowe's works supports the influence of Marlowe or even possible collaborations with Marlowe during this early phase of Shakespeare's literary life.

From a technical point of view, the word rank order and frequency method applied here complements traditional approaches to the Shakespeare-Marlowe authorship debate. The method was originally developed for studying a wide range of symbolic sequences[26,27], and, therefore, assumes minimal knowledge about any specific language. Our new method is based on a simple assumption, namely that different authors have a preference for certain words, which they use with higher frequency. In our analysis, synonyms, homonyms, plural usages, and verbs with different tenses are all considered as different words (types). We propose that those variants also represent a "selection" by the author, and therefore should not be treated as the same word. We observed that both Shakespeare and Marlowe evolved their distinct "preferences" for word usages over time. The nature of this phenomenon in genetic (DNA) sequences is related to selection pressures attributable to factors such as nutrition, illness, or other environmental influences. In written languages, this phenomenon is possibly related to deep levels of cognitive function, modulated by increasing life experience and knowledge. The behaviour of word usages of the types under study with this method is unlikely to be consciously manipulated by the author.

SUMMARY

Using a new quantitative approach to information categorization -- the word rank order and frequency method, we have revisited the Shakespeare-Marlowe authorship controversy. Our findings, using this validated and fully objective computer-based method, indicate the following three key results.

(1) The major dramatic works attributed to William Shakespeare are clearly distinct from

those of Christopher Marlowe. Our analysis does not address the question of whether "Shakespeare" was the Stratford dramatist, but does indicate that the author of the body of Shakespearean plays is not Marlowe.

(2) However, one play recently restored to the Shakespearean canon, *Edward III*, is more likely to have been predominantly composed by Marlowe than by Shakespeare. Based on these new and compelling statistical findings, the "canonization" of the play should be reconsidered.

(3) The early Shakespeare histories, while distinct from Marlowe, are closest to the Canterbury dramatist's works in style, suggesting that Marlowe importantly influenced the young Shakespeare. The later plays demonstrate a stylistic evolution that makes these dramas markedly distinct from Marlowe's corpus.

Acknowledgements

We gratefully acknowledge the support from the U.S. National Center for Research Resources of the National Institutes of Health (NIH) (P41-RR13622), the NIH/National Institute on Aging (P60-AG08812), and the G. Harold and Leila Y. Mathers Charitable Foundation.

VIII. Appendix

In this Appendix we briefly outline the mathematical formula developed to measure the "dis-similarity" between two texts. As previously published[26] and described in the main text, the "distance" (or dissimilarity) between any two texts can be quantified by measuring the scatter of these points (see Fig 2.) from the diagonal line in the rank order comparison plot. Greater distance indicates less similarity and vice versa. Therefore, we can define the distance (D) between two texts, T_1 and T_2, as

$$D(T_1,T_2)=\frac{1}{N_{12}}\sum_{k=1}^{N_{12}}\left|R_1(w_k)-R_2(w_k)\right|F(w_k). \qquad (1)$$

Here $R_1(w_k)$ and $R_2(w_k)$ represent the rank of a specific word, w_k, in texts T_1 and T_2, respectively. N_{12} is the number of total shared words used in texts T_1 and T_2. The absolute difference of ranks, $\left|R_1(w_k)-R_2(w_k)\right|$, is proportional to the euclidean distance from a scattered point to the diagonal line. This term is then multiplied by a weighting function, $F(w_k)$, to take into account that not all points on the rank order comparison plot are equal. The more frequently used words should contribute more to defining the style of an author.

We select the weighting function $F(w_k)$ to be the sum of Shannon's entropy[54] for w_k in texts T_1 and T_2.

$$F(w_k)=\left[-p_1(w_k)\log(p_1(w_k))-p_2(w_k)\log(p_2(w_k))\right]/Z$$

where Z is the normalization factor such that $\sum_{k=1}^{N_{12}} F(w_k) = 1$. Here $p_1(w_k)$, and $p_2(w_k)$ represent the probability of a specific word, w_k, in texts T_1 and T_2, respectively. The selection of Shannon's entropy to be the weighting function is to ensure that words occurring with higher probability will be more heavily weighted.

Reference List

1. Zeigler, W.G. It Was Marlowe: A Story of the Secret of Three Centuries. Chicago (1895).

2. Hoffman, C. The Murder of the Man Who Was Shakespeare. Julian Messner, Inc., New York (1955).

3. Bloom, H. Shakespeare: The Invention of the Human. Riverhead Books, New York (1998).

4. Bate, J. The Genius of Shakespeare. Picador, London (1997).

5. Garber, M. Shakespeare's Ghost Writers: Literature as Uncanny Causality. Methuen, New York (1987).

6. Erne, L. Shakespeare's 'Edward III': An early play restored to the canon. Archiv fur das Studium der Neueren Sprachen und Literaturen 236, 425-427 (1999).

7. Proudfoot, R. The Reign of King Edward the Third (1596) and Shakespeare. Proceedings of the British Academy 71, 169-185 (1985).

8. Sams, E. Shakespeare's Edward III. Yale University Press, New Haven (1996).

9. Wentersdorf, K. The Authorship of Edward III. Ph.D. Thesis, University of Cincinnati. (1960).

10. Robertson, J.M. Did Shakespeare Write Titus Andronicus? AMS Press, New York (1905).

11. Mendenhall, T. The characteristic curves of composition. Science 11, 237-249 (1887).

12. Yule, G.U. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authorship. Biometrika 30, 363-390 (1938).

13. Morton, A.Q. The authorship of Greek prose. J. Roy. Stat. Soc. A. 128, 169-233 (1965).

14. Mosteller, F. & Wallace, D.L. Inference in authorship problem - a comparative-study of discrimination methods applied to authorship of disputed Federalist papers. J. Am. Stat. Assoc. 58, 275-309 (1963).

15. Cooke, J.E.E. The Federalist. Word Publishing Company, Meridian Books, Cleveland, Ohio (1961).

16. Burrows, J. Questions of authorship: attribution and beyond - a lecture delivered on the occasion of the Roberto Busa Award ACH-ALLC 2001, New York. Comput. Humanit. 37, 5-32 (2003).

17. Holmes, D.I. Authorship attribution. Comput. Humanit. 28, 87-106 (1994).

18. Tweedie, F.J. & Baayen, R.H. How variable may a constant be? measures of lexical richness in perspective. Comput. Humanit. 32, 323-352 (1998).

19. Merriam, T. & Matthews, R. Neural computation in stylometry II: an application to the works of Shakespeare and Marlowe. Lit. Linguist. Comput. 9, 1-6 (1994).

20. Havlin, S. The distance between Zipf plots. Physica A 216, 148-150 (1995).

21. Efron, B. & Thisted, R. Estimating number of unseen species - how many words did Shakespeare know. Biometrika 63, 435-447 (1976).

22. Ellis, B.G. & Dick, S.J. “Who was shadow?” - The computer knows: applying grammar-program statistics in content analysis to solve mysteries about authorship. Journalism & Mass Communication Quarterly 73, 947-962 (1996).

23. Burrows, J. 'Delta': a measure of stylistic difference and a guide to likely authorship. Lit. Linguist. Comput. 17(3), 267-287 2002.

24. Hoover, D.L. Frequent word sequences and statistical stylistics. Lit. Linguist. Comput. 17, 157-180 (2002).

25. Mosteller, F. & Wallace, D.L. Applied Bayesian and Classical Inference: The Case of the Federalist papers. Springer-Verlag, New York (1984).

26. Yang, A.C.C., Peng, C.K., Yien, H.W. & Goldberger, A.L. Information categorization approach to literary authorship disputes. Physica A, in press. (2003).

27. Yang, A.C.C., Hseu, S.S., Yien, H.W., Goldberger, A.L. & Peng, C.K. Linguistic analysis of the human heartbeat: using frequency and rank order statistics. Phys. Rev. Lett. 90:108103 (2003).

28. Darwin, C. On the Origin of Species by Means of Natural Selection. John Murray, London (1859).

29. Watson, J.D. & Crick, F.H.C. Molecular structure of nucleic acids - a structure for deoxyribose nucleic acid. Nature 171, 737-738 (1953).

30. Felsenstein, J. PHYLIP (Phylogeny Inference Package). (3.5c). 1993. Department of Genetics, University of Washington, Seattle.

31. Saitou, N. & Nei, M. The neighbor-joining method - a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406-425 (1987).

32. Page, R.D.M. & Holmes, E.C. Molecular Evolution: A Phylogenetic Approach. Blackwell Science Inc., Cambridge (1998).

33. Vilensky, B. Can analysis of word frequency distinguish between writings of different authors? Physica A 231, 705-711 (1996).

34. Constitution Society. Federalist Papers (http://www.constitution.org).

35. Merriam, T. An experiment with the Federalist Papers. Comput. Humanit. 23, 251-254 (1989).

36. Tweedie, F.J., Singh, S. & Holmes, D.I. Neural network applications in stylometry: The Federalist Papers. Comput. Humanit. 30, 1-10 (1996).

37. Yuan Ze University. Chinese website: The Dream of the Red Chamber (http://cls.admin.yzu.edu.tw/hlm/).

38. Hu, S. et al. Collected Chinese papers: Textual research of The Dream of the Red Chamber. The Far East Book Comapny, Taipei, Taiwan (1985).

39. Literature Online (Lion) (http://www.chadwyck.com/products/pt-product-Lion.shtml).

40. Virginia Electronic Text Center (http://etext.lib.virginia.edu/).

41. The Complete Works of William Shakespeare (http://the-tech.mit.edu/Shakespeare/).

42. The Complete Works of Christopher Marlowe (http://www.perseus.tufts.edu/Texts/Marlowe.html).

43. Godshalk, W.L. Dating Edward III. Notes and Queries 42, 299-300 (1995).

44. Capell, E. Prolusions. J. and R. Tonson, London (1760).

45. Bate, J. Shakespeare's 'Edward III', an early play restored to the canon - Sams, E. The Times Literary Supplement 3-4 (1997).

46. Nielson, C.T. Shakespeare's 'Edward III': an early play restored to the canon. Renaissance Quarterly 51, 1039-1040 (1998).

47. Slater, E. The Problem of The Reign of King Edward III: A Statistical Approach. Cambridge University Press, Cambridge (1988).

48. Merriam, T. Shakespeare's 'Edward III' - Sams, E. Notes and Queries 44, 261-262 (1997).

49. Merriam, T. Edward III. Lit. Linguist. Comput. 15, 157-186 (2000).

50. Merriam, T. Marlowe's hand in Edward III revisited. Lit. Linguist. Comput. 11(1), 19-22 (1996).

51. Wentersdorf, K. The Date of Edward III. Shakespeare Quarterly 16, 227-231 (1965).

52. Greene, R. Perimedes the Blacksmith. London (1588).

53. Wells, S. & Taylor, G. William Shakespeare, The Complete Works. Clarendon Press, Oxford (1988).

54. Shannon, C.E. A mathematical theory of communication. Bell. Labs. Tech. 27, 379-423 (1948).

a. The Winter's Tale (William Shakespeare)

Word	Rank	Frequency
The	1	857
I	2	701
And	3	657
To	4	635
Of	5	475
You	6	472
A	7	419
My	8	405
That	9	338
Not	10	305
...	...	...

Total different words: 3703

b. Cymbeline (William Shakespeare)

Word	Rank	Frequency
The	1	966
I	2	771
And	3	713
To	4	671
Of	5	525
A	6	459
You	7	424
My	8	383
That	9	381
In	10	320
...	...	...

Total different words: 4042

c. Bonduca (John Fletcher)

Word	Rank	Frequency
And	1	667
The	2	584
I	3	552
To	4	432
A	5	393
You	6	286
Of	7	273
Petillius	8	224
Your	9	214
That	10	203
...	...	...

Total different words: 3124

Figure 1. Word rank order and frequency lists for **a**, *The Winter's Tale*; **b**, *Cymbeline*; and **c**, *Bonduca*. To construct the list for each play, we count the occurrences of each word, and then sort them by descending frequency. For example, the first ranked word in *The Winter's Tale* is "the", followed by the second ranked word "I", and so on. The resulting rank frequency list represents the statistical hierarchy of word usage of the original text.

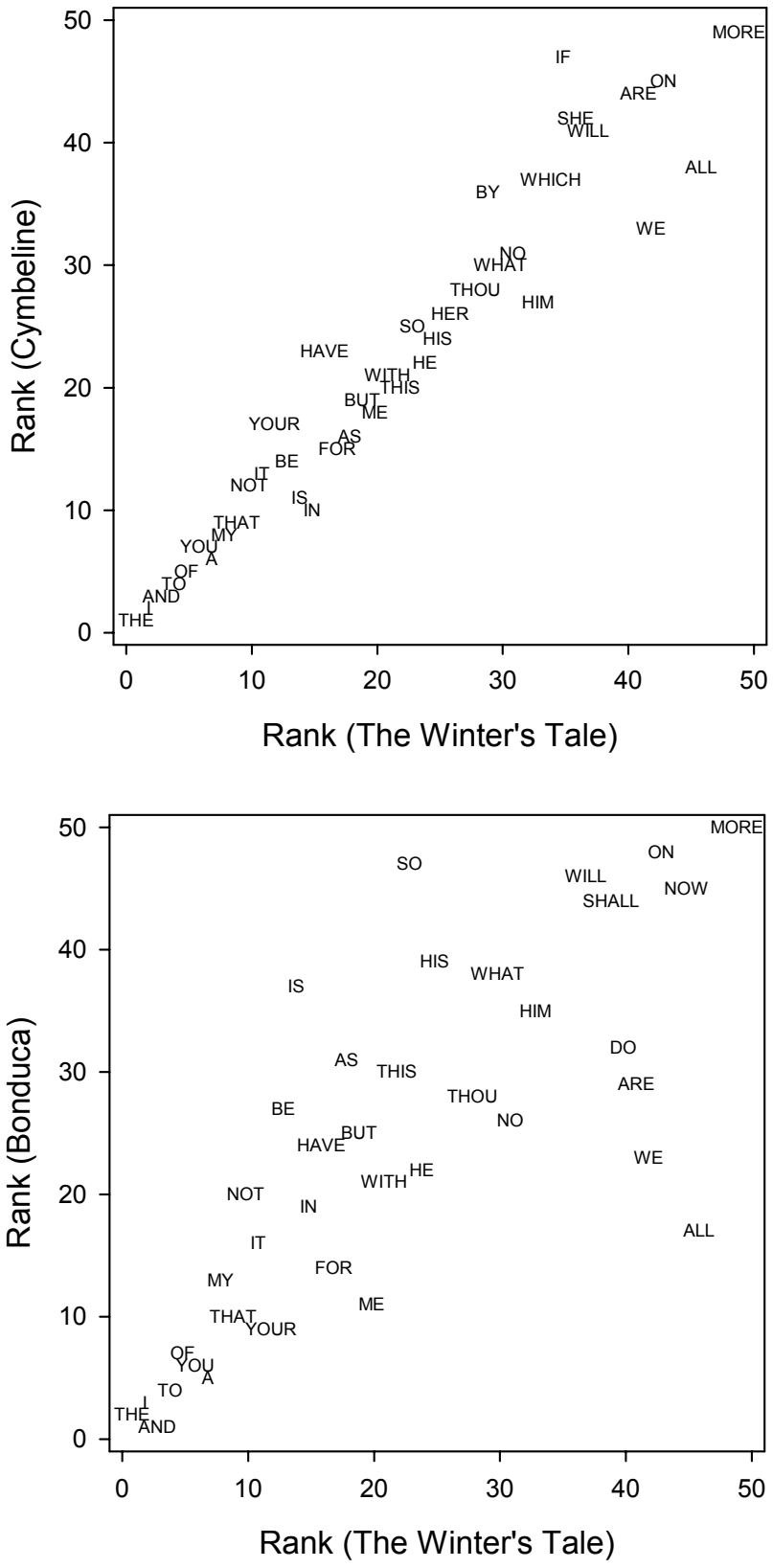
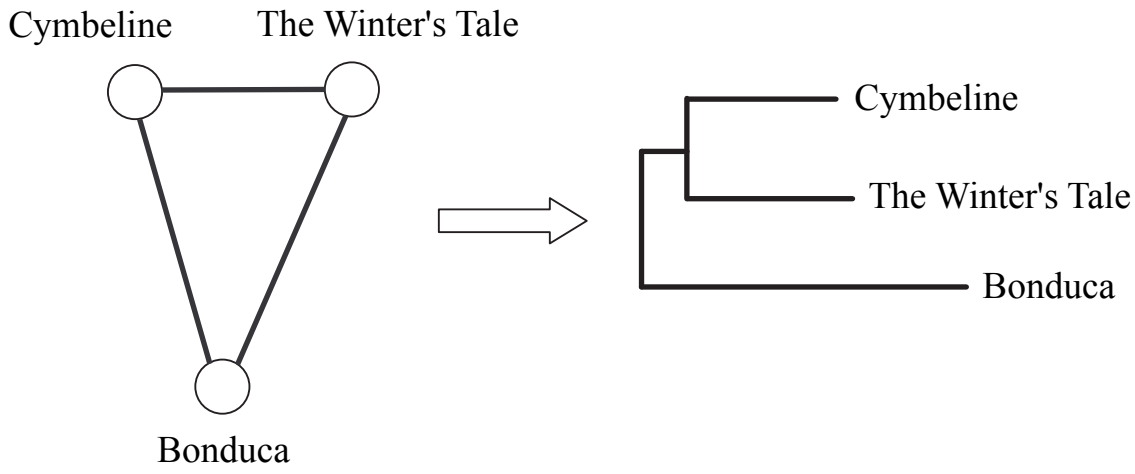


Figure 2. Rank order comparison of the top ranking words for **a**, two of Shakespeare's plays: *The Winter's Tale* versus *Cymbeline*; and **b**, *The Winter's Tale* versus John Fletcher's *Bonduca*. Words from the two Shakespeare plays fall close to the diagonal, indicating nearly identical ranking, in contrast to the Fletcher-Shakespeare comparison.

a. Schematic illustration of phylogenetic analysis



b. Phylogenetic tree of Shakespeare-Fletcher databases

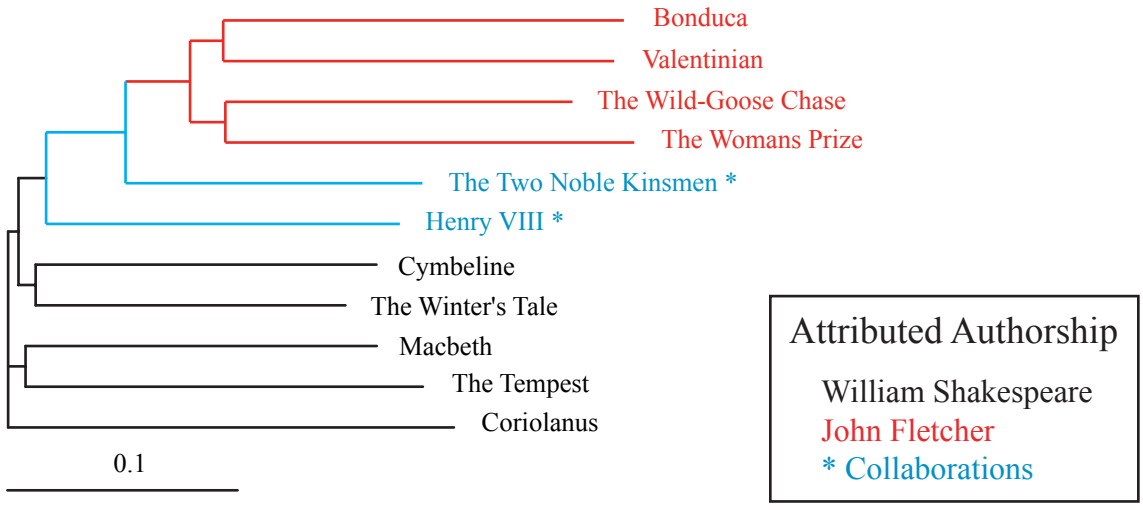


Figure 3. Schematic illustration of phylogenetic analysis. **a**, The triangular relationship of three plays determined by the word rank order and frequency method. The two Shakespeare plays, *The Winter's Tale* and *Cymbeline*, are closest to each other, and relatively distant from the Fletcher play, *Bonduca*. The resulting phylogenetic tree (right) can be estimated from such triangular relationships. **b**, A mini-database containing five of Shakespeare's late plays, four of John Fletcher's sole-authorship plays, as well as two of their putative collaborations, *Henry VIII* and *The Two Noble Kinsmen*.

Authorship	Title	Word Counts
L. F. Baum	The Wonderful Wizard of Oz	39282
	The Marvelous Land of Oz	41571
E. Brontë	Wuthering Heights	116534
L. Carroll	Alice's Adventures in Wonderland	26505
	Through the Looking-glass and What Alice found there	29053
A. Conan Doyle	The Sign of Four	43125
	The Hound of the Baskervilles	59233
	The Valley of Fear	57746
H. James	Confidence	76512
	The Europeans	59800
St Luke	Gospel according to St Luke	25939
	The Acts of the Apostles	24246
J. London	The Sea Wolf	105925
	The Call of the Wild	31891
H. G. Wells	The War of the Worlds	60187
	The Invisible Man	48599

Data from Oxford Text Archive

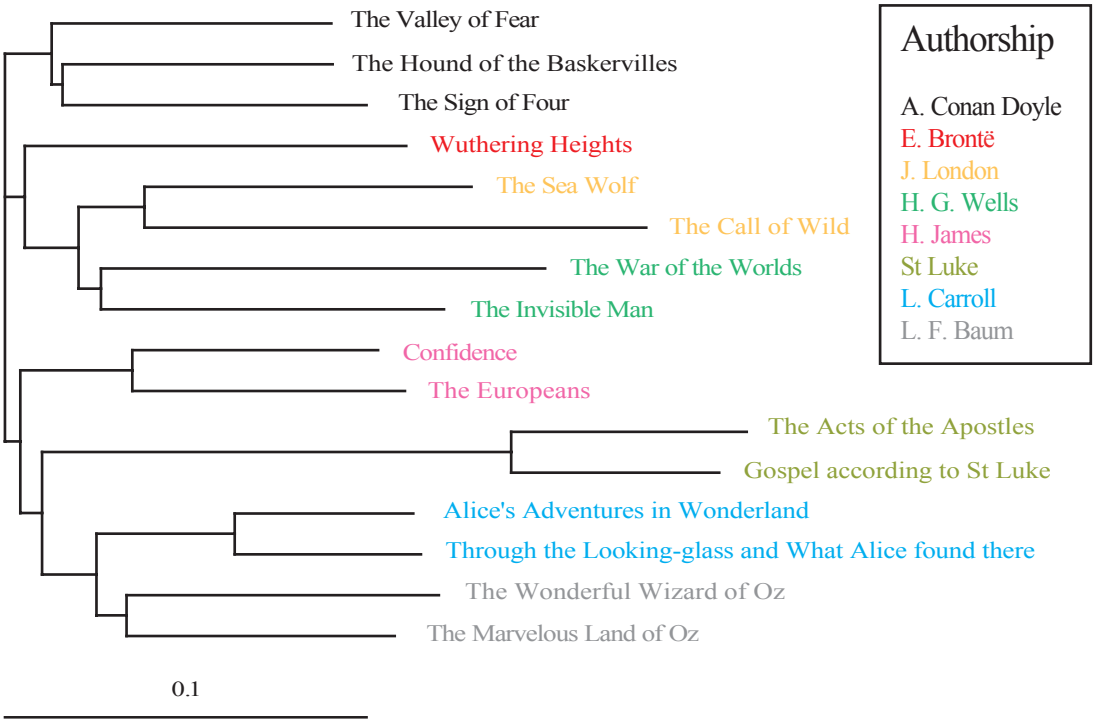
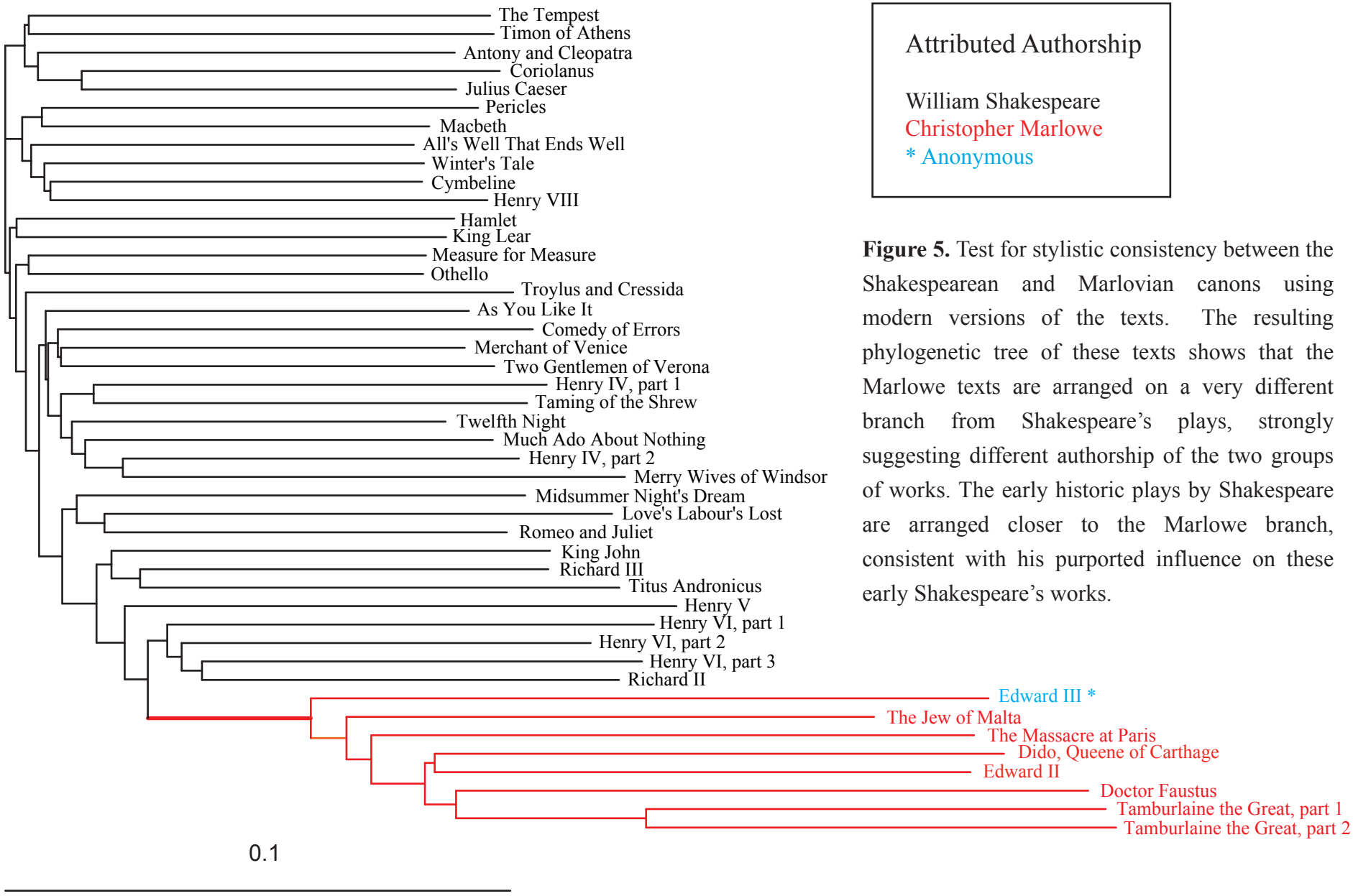


Figure 4. Initial validation of the word rank order and frequency method (see text).



Attributed Authorship

William Shakespeare
Christopher Marlowe
* Anonymous

Figure 5. Test for stylistic consistency between the Shakespearean and Marlovian canons using modern versions of the texts. The resulting phylogenetic tree of these texts shows that the Marlowe texts are arranged on a very different branch from Shakespeare's plays, strongly suggesting different authorship of the two groups of works. The early historic plays by Shakespeare are arranged closer to the Marlowe branch, consistent with his purported influence on these early Shakespeare's works.

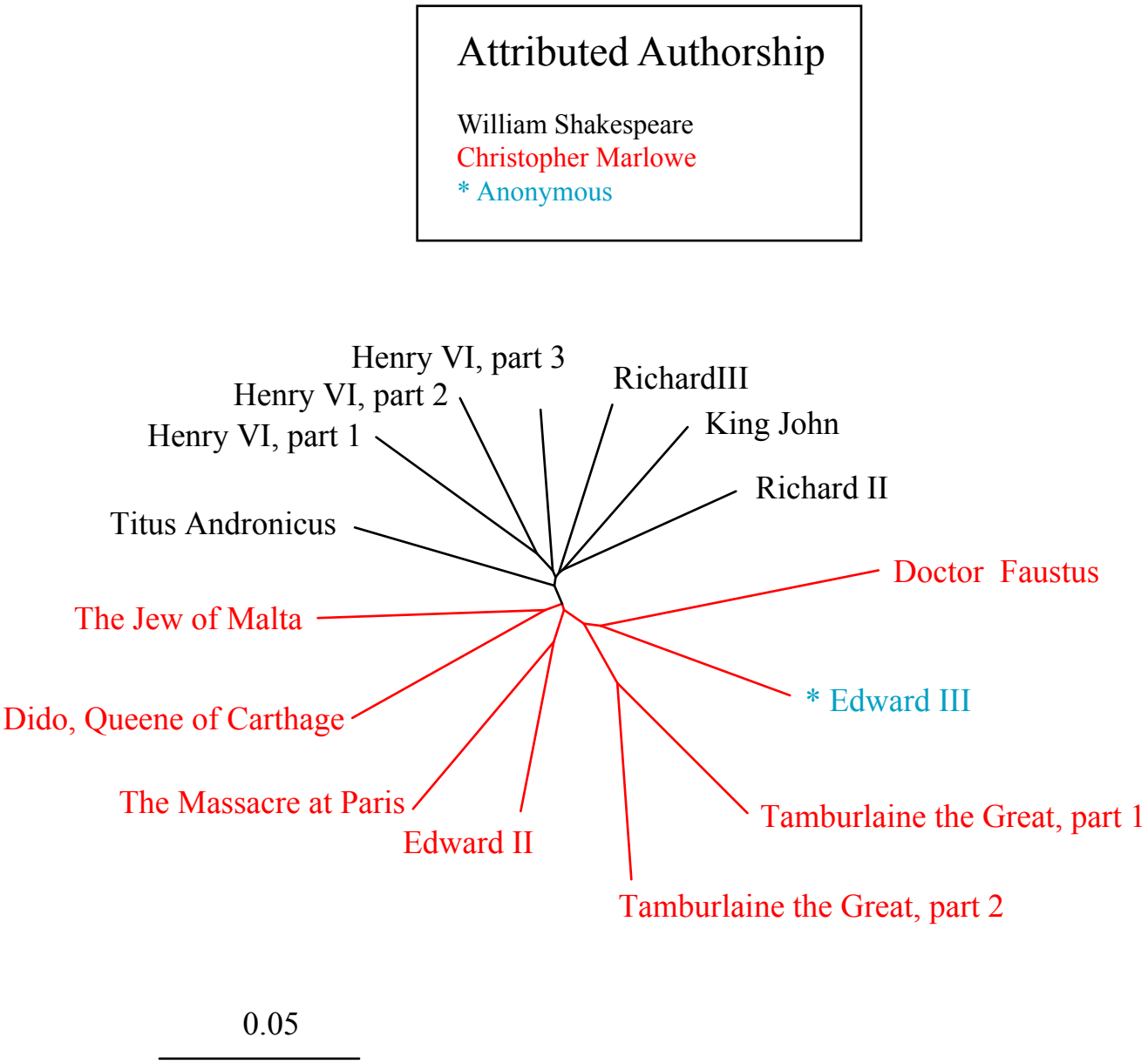


Figure 6. Test for the stylistic consistency between the Shakespearean and Marlovian canons using the old versions of the texts. The distance between the early Shakespeare and Marlowe texts is even closer than that in Figure 5.

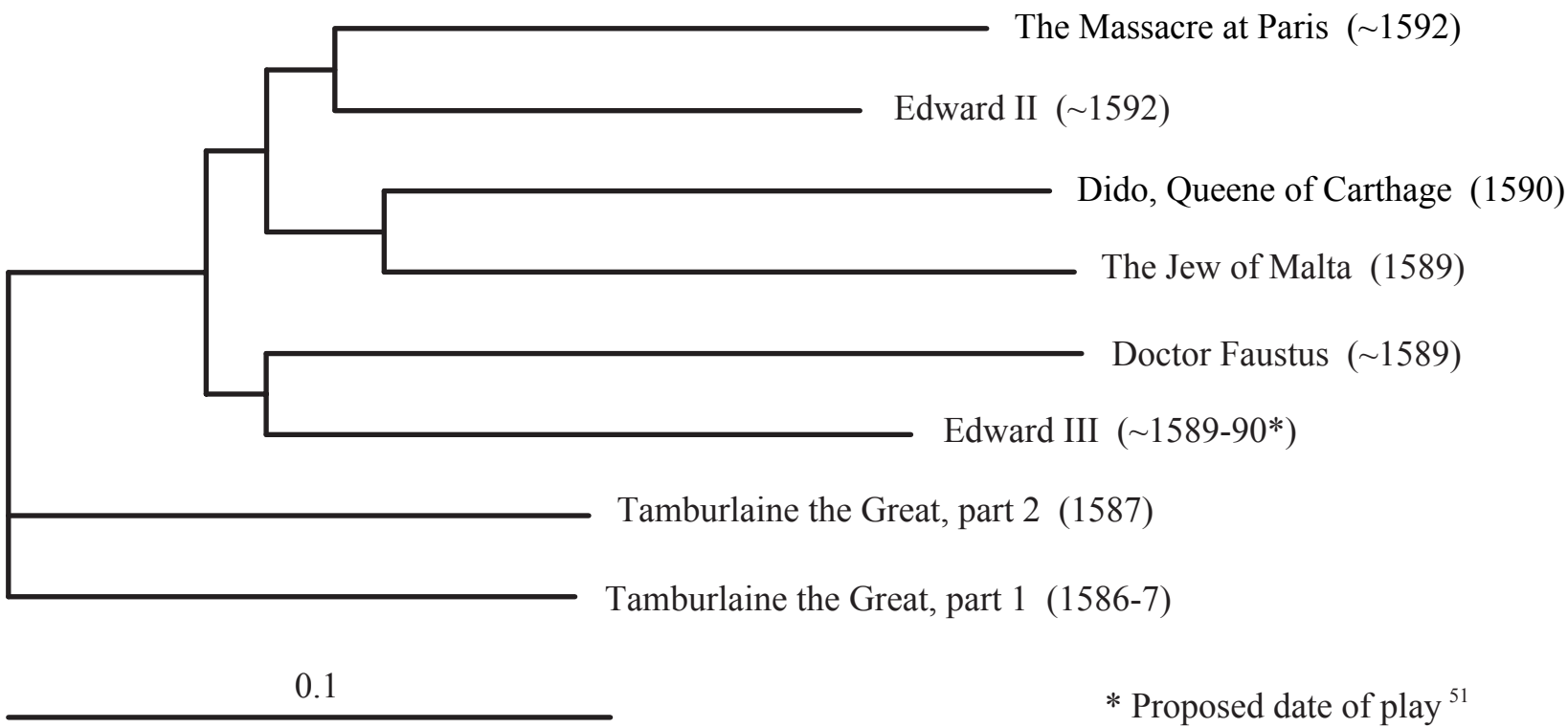


Figure 7. Evolutionary tree of Marlowe plays. We assign the root of the tree (the very first play by Marlowe) as *Tamburlaine the Great*, dated around 1587[52]. *Edward III* falls between *Tamburlaine the Great, Part 2* (1588) and *Dr. Faustus* (1589-90), suggesting the former play dates from around 1588-90.

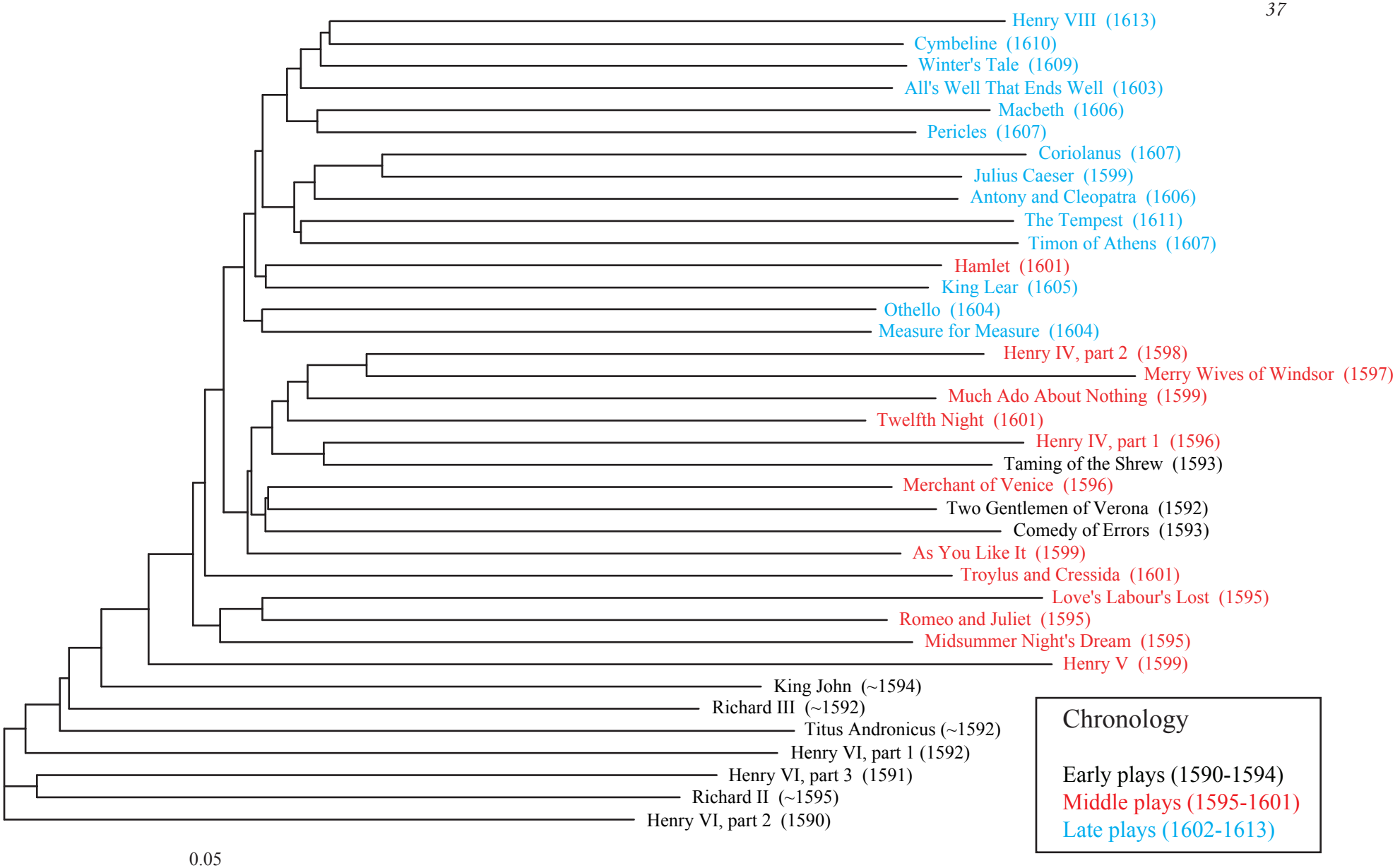


Figure 8. Evolutionary tree of Shakespeare plays. We assign the root (first play) as the second part of the *Henry VI* series as proposed by Wells and Taylor[53]. The resulting Shakespeare dramatic phylogeny reasonably places the early plays (1590-94), middle plays (1595-1601), and late plays (1602-1613) from the bottom to the top of the tree.