Using phylogenetic methods to explore culinary innovations in culture

# Introduction

The ability of phylogenetic methods to build maps of relatedness based on morphological states an be exploited to understand the transmission of cultural knowledge. In the research following, phylogenetic methods were tested first to see if they had power to accurately cluster data on culinary innovations found in recipes to their country of origin. The clustering of various phylogenetic methods were subjectively assessed to select the best method. The resultant best clustering was then used to assess potential relations between food.

While using phylogenetic methods to test relatedness between cultural aspects is not a new idea as shown between building language trees and music trees (cite language study), the full potential of this concept has yet to be exploited. Cultural innovations are passed down through generations within a populace. Cultures have been able to maintain relative development in isolation bar their near neighbors until the invention of modern modes of travel. Horizontal transmission of cultural elements have been relatively slow even during the age of ship travel. As the world becomes more connected through air travel by the use of airplane, and we continue to move towards a global economy, cultures are beginning to merge into a continuum. It can be expected that in the near future, cultural elements may become merged to the point where no distinct signal of cultural origin can be distinguished with it extensive look to the past, highlighting the importance of this work.

# Methods

## Assembling the dataset

The top 20 recipes from a set of countries/regions were scrapped from AllRecipes.com using a public script that had been repurposed by Dr. Jennifer Chang. The data scraped from the page included recipe country/region, recipe id, recipe title, and recipe ingredients. The ingredients were used as the characters for phylogenetic reconstruction. This data was further processed to be in a format where the characters were presented in a binary manner. This file was processed into tnt, phylip, and nexus format for use with different tools downstream in the analysis. In total, 356 recipes were used, composed of a total of 223 ingredients.

## Assessing Treeness

The treeness of the data should have been assessed before attempting any work, retrospectively. This would have given a good indication if this route was worth pursuing.

## Inferring relatedness by parsimony methods

The characters of each recipe were defined as discrete and binary. TNT was used to analyze the characters (see recipes.tnt in the data folder). TNT was set to hold 1000 trees for swapping, and performed random addition of sequences with tree rearrangements in 200 replicates. A majority consensus tree was attempted to be built, though was of low quality containing many polytomies. A single tree with the best score (1712) was used for comparison.

tnt p renamedfood.txt, log highep.out, rep+1, hold 1000, mult=replic 200, le, resample,majority, taxname=,export - hitree.tre

## Distance methods

Pairwise distances between recipes were calculated by summing the results of an exclusive or (XOR) between binary characters. If one recipe had an ingredient that was not present in the compared recipe, the distance between recipes was incremented by one. Trees were built using neighbor-joining methods, made available through the APE package in R. Code is available in the repository (njTree.R).

## Maximum likelihood methods

Maximum likelihood trees were built using RAxML under a general time reversible (GTR) model for the multi-state morphological data.

raxmlHPC-PTHREADS-SSE3 -p 777 -m MULTIGAMMA -s test.phy -n maxlikefood

## Bayesian methods

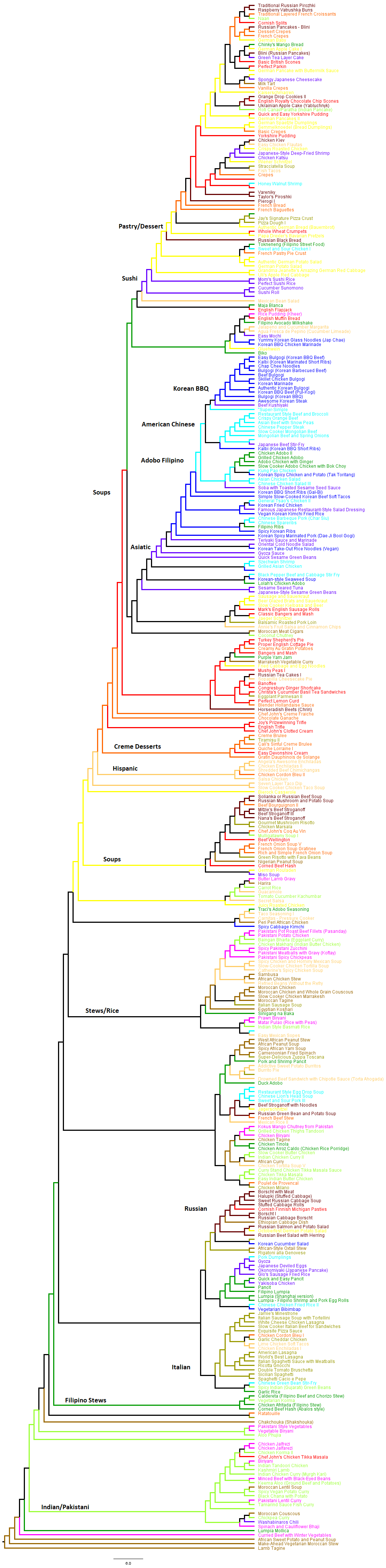
Bayesian analysis were run on the hpc cluster using mrbayes. The likelihood coding model was to variable sampling due to morphological features of variance being selected (ingredients per recipe). The Dirichlet prior was set to infinity to assume all character states had equal frequency, and the site specific rate models was set to variable, although the data had a single partition of, which the program would have assumed an average rate of substitution being 1. After 10,000,000 generations sampling every 1000 trees, runs on two chains, the analyses failed to converge. There was an observed lack of swapping between hot and cold chains, potentially contributing to the failure to converge. Single high scoring trees in the upper part of tree space were comparable to the maximum likelihood trees and were used for comparison. Files related to the bayesian run are available in the repository.

# Results

## Parsimony Methods

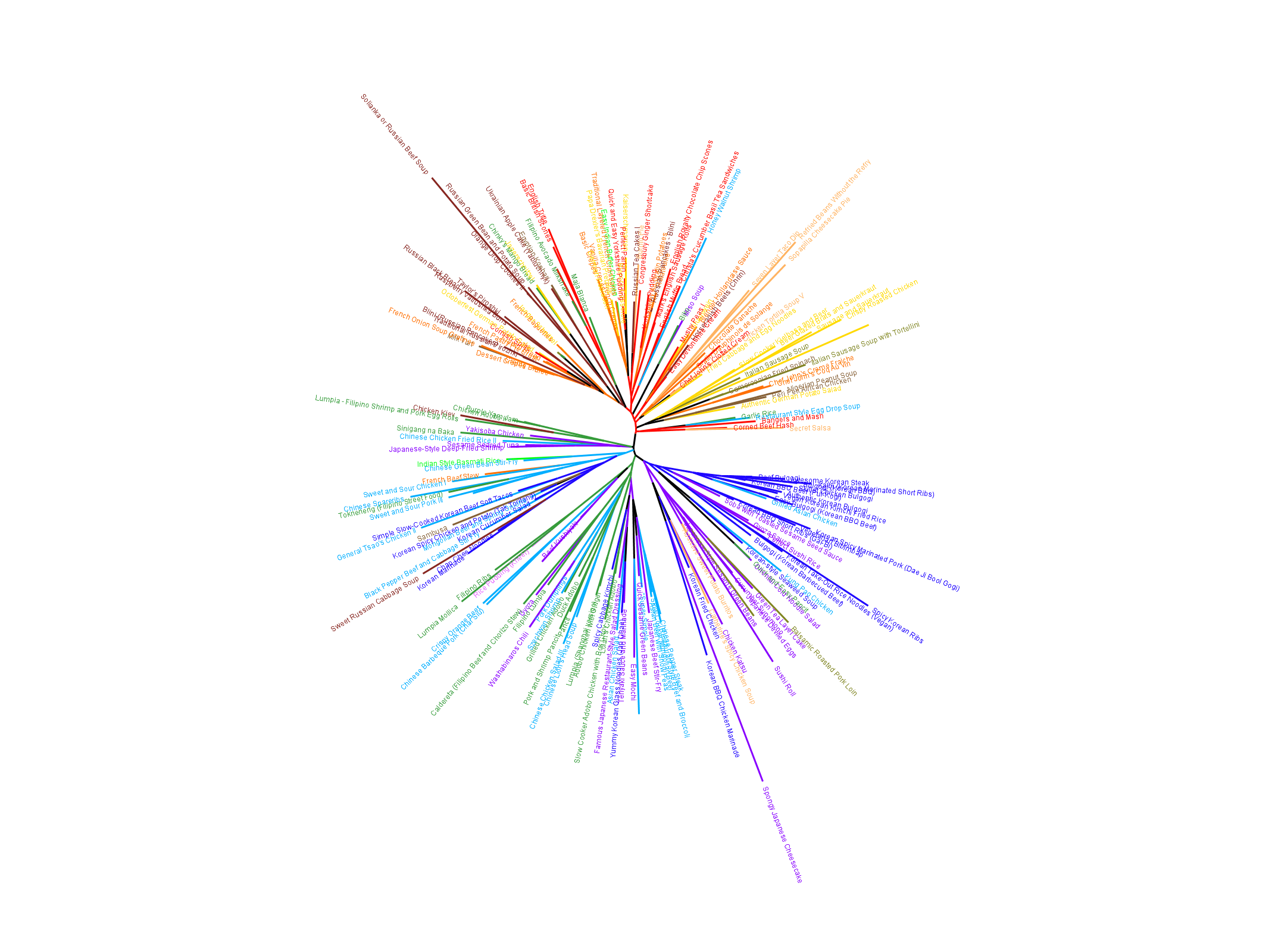
The author was unable to create a sensible consensus using parsimony methods. Inspection of a single tree generated from parsimony methods shows weak clustering based on country of origin, with repeated motifs. The tree is rooted to the recipe “Lamb Tagine” from Africa. European countries tend to cluster together with Germanic foods proceeding French foods proceeding English foods in terms of the deepest most recent common ancestor (MRCA). While it would be incorrect to say Germanic foods are ancestral to British foods as both the recipes are leafs still in existence today, it does emphasize how closely these three countries share culinary traditions. South-east Asian countries tended to cluster together in terms of recipes. India and Pakistan closely clustered, and oddly Mexico and African recipes tended to cluster into these groups. Italy formed its own distinct cluster on the parsimony tree.

The repeated motifs on the tree appear to be caused by clustering by recipe type. The German-French-British motif appears in a region of the tree dominated by desserts/pastries, then again in a soup dominant part of the tree. There was also a pasta dominant, a stew dominant, and a BBQ dominant section of the tree as well. The parsimony methods likely caused this due to a different set of ingredients being needed bake rather than broil.



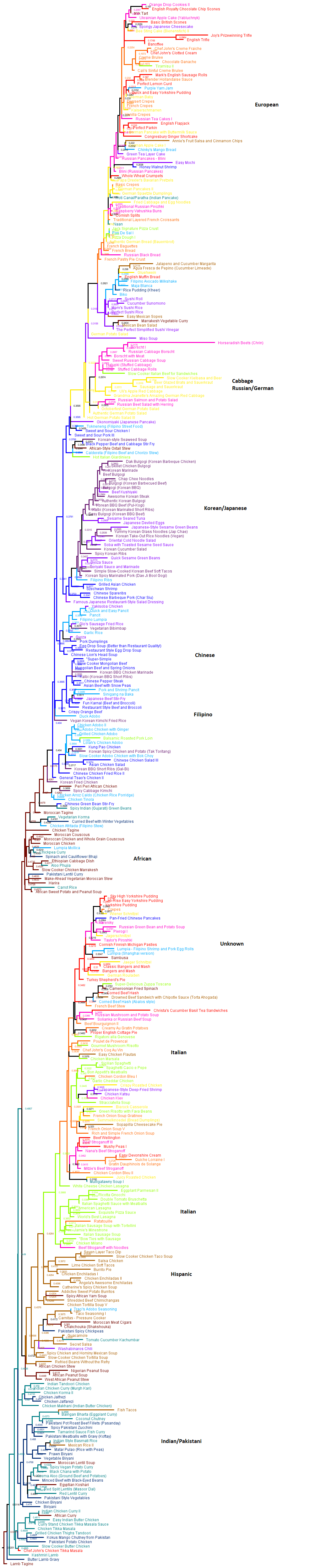
## Distance Methods

A user error caused a portion of the data to not be used for building the neighbor joining tree. The tree was rooted to “Nigerian Peanut Soup” because the author failed to realize taxa was missing from the tree (Lamb Tagine) until the last moment. The neighbor joining method showed a clear distinction between Asiatic recipes and all other (European, African, Hispanic) when viewed as an unrooted tree. There was better clustering on a per country level, with more contiguous grouping of clades. The “honey walnut shrimp” Chinese recipe, Chinese egg drop soup, as well as a few Filipino recipes and Indian buttered chicken clustered into the European side of the tree unexpectedly, suggesting either a type of convergence in the culinary tradition, or an inauthenticity in the recipe. The methods used for building the neighbor joining tree did not take into account the types of ingredients used directly perhaps allowing better broad clustering than parsimony methods.



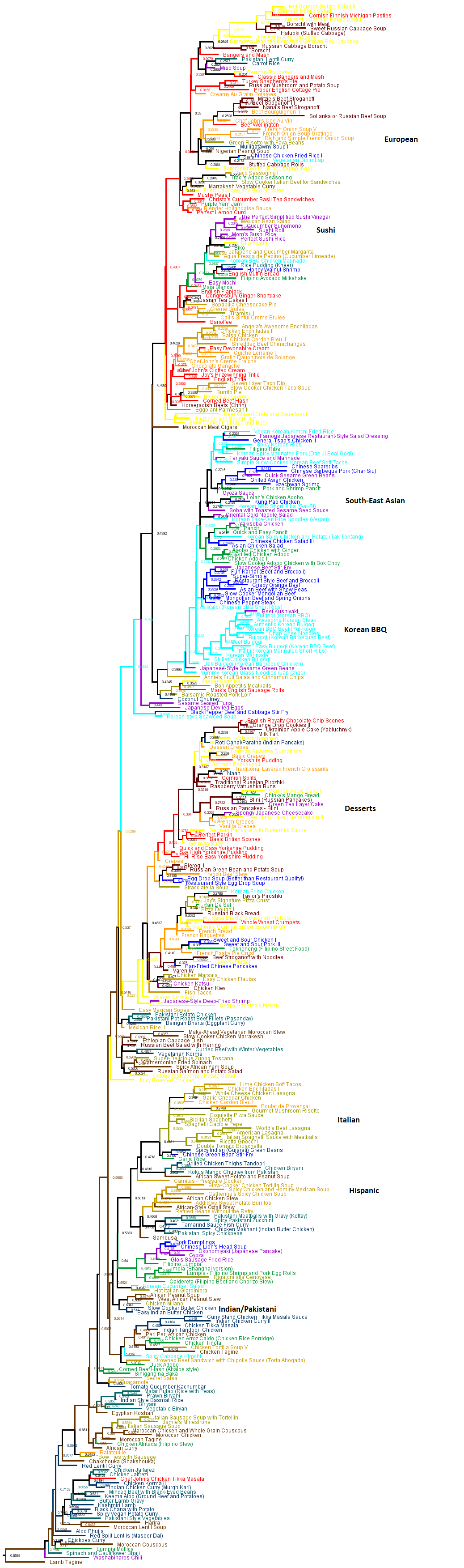
## Maximum likelihood methods

The tree is rooted to the recipe “Lamb Tagine” from Africa. Maximum likelihood methods show good ability to separate out recipes by culinary culture, creating multiple contiguous groupings. It further is able to show shared culinary tradition, such as between Pakistan and Indian, as well as between Korea and Japan, and to a lesser extent China. Despite the improvement of clustering form prior methods, the bootstrap values are low, suggesting a less robust support for the results shown.



## Bayesian Methods

The tree is rooted to the recipe “Lamb Tagine” from Africa. The Bayesian analysis failed to converge and create a reasonable consensus tree, so a tree was selected from the higher scoring regions of tree space. The result is comparable with the maximum likelihood tree, albeit a little less straight forward. There is obvious clustering between south-east Asian cultures as well as between India and Pakistan. Many of the recipes labelled as “Mexican” are grouped inside the European parts of the tree. There are still low support values for the branches, emphasizing the uncertainty of the placement. Future runs should consider longer searches through tree space with parameters set that would force swapping between chains, to avoid becoming stuck in any area of tree space.



# Discussion

Before any evaluation of the methods or their results can be had, it is imperative to discuss the caveats and limitations of this experimental exercise of phylogenetic methods. First and foremost, AllRecipes.com is a low-quality data source for recipes. This is exemplified by some recipes calling for premade ingredients, such as taco mix or canned soups, which are much more readily available in the United States rather than globally. The majority of recipes present on AllRecipes.com are interpretations of authentic recipes, most likely from an American perspective. Because the recipes are not being made in traditional ways from traditional ingredients, the signal is obscured. Better reconstruction and clustering of recipe groups may have been found if authentic recipe books or other likewise stronger data source had been used.

Additionally, upon closer inspection of the data, repeats in recipes can be seen. While the recipes selected were listed as the best of group on AllRecipes.com, there are repeats of recipes such as bulgogi. While not inherently bad for the phylogenetic process, it gives an artificial sense of success in clustering using the phylogenetic method.

For this study, low information content terms were used. Rather than stick with higher content terms such as “yellow onion”, “green onion”, etc., the author opted to truncate the term to simply onion, potentially losing information that could have helped better discriminate between different classifications better. To sum up the prior three points simply, attempting to put “garbage in” into phylogenetic methods will lead to getting “garbage out.” Careful data selection and curation is essential to getting meaningful results using phylogenetic methods, whether it is from morphological data from recipes, or species data from animals.

From all the methods, maximum-likelihood produced the most contiguous tree that was consistent with prior knowledge on how food practices spread based on proximity. Parsimony methods over-fragmented the population of recipes with respect to ingredient, forming clades more consistent with the type of food being served. Distance methods could sort between Asiatic and European dishes, but were not able to cluster very well based on country. This is potentially because of a loss in data on what the characters were. Maximum-likelihood results had the best clustering per known countries. Bayesian methods did not converge, and should be repeated in the future, but the higher-ranking trees are near comparable to the maximum likelihood trees.

From the maximum likelihood tree, it is difficult to discern a distinct origin between Pakistani dishes and Indian dishes, which is consistent with the known culture of these two countries. They had split in 1947 after World War II, a decision made by the British. The South-East Asian countries of China, Japan, Korea, and the Philippines clustered together in terms of recipes, as would be expected with prior knowledge on trade and conflict between these countries. Despite the similarities shown on the tree, each of these cultures has a unique culinary tradition, which may have been missed during clustering due to the use of low quality recipes or low information content terms. Interestingly, some Hispanic recipes are rooted inside European recipes, while another set are located near Indian/Pakistani dishes. Further investigation could be made to see if this is due to a heavy European influence on the culinary tradition due to colonization, or if this is an artifact of the low quality dataset. English, French, German, and Russian recipes are all tightly mixed, suggesting perhaps heavy mixing of culinary innovation. This would be consistent with knowledge of past trade and conflict. Perhaps interestingly, it should be noted that two Japanese dishes are nested inside the European dessert section of the tree, “spongy Japanese cheesecake” and “green tea layered cake”. Prior knowledge will show the innovation for the creation of these types of cake are of European origin (Italian Chef Giovan Battista Cabona working in Giovan Battista Cabonahe Spanish court (cite)), with the Japanese adopting the culinary innovation and modifying it to suite their palate in the recent past through a horizontal transmission event of culture. Other similar events can be found in the maximum likelihood tree, but a subject matter expert would be required to appropriately interpret the result.

With modern modes of transportation and the innovation of the internet, cultures are able to horizontally integrate at a much faster rate than ever before. It is increasing becoming difficult to find cultures in isolation with unique hallmarks. Despite globalization, many countries still maintain culinary traditions which were developed based on regional availability as well as locally defined palettes. Globalization is spreading ‘invasive species’ of food around the world, some which sustain very well.

# References