Applying population genetic approaches within languages:
Finnish dialects as linguistic populations

Kaj Syrjänen[1*], Terhi Honkola[2*], Jyri Lehtinen[3], Antti Leino[4], Outi Vesakoski[5]
* Shared first authorship

[1] School of Language, Translation and Literary studies
FI-33014 University of Tampere
kaj.jaakko.syrjanen@uta.fi

[2] Department of Biology,
FI-20014 University of Turku
terhi.honkola@utu.fi

[3] Department of Modern Languages,
FI-00014 University of Helsinki
jyri.lehtinen@helsinki.fi

[4] School of Language, Translation and Literary studies
FI-33014 University of Tampere
unni-paiva.leino@uta.fi

[5] Department of Biology,
FI-20014 University of Turku
outi.vesakoski@utu.fi

Abstract

The adoption of evolutionary approaches to study language change as a type of non-biological evolution has gained increasing interest and brought variety of quantitative tools for linguistics. The focus has thus far mainly been on language families, or 'linguistic macroevolution' and have taken the shape of linguistic phylogenetics. Here we explore whether evolutionary methods could be applicable for studying intra-lingual variation ('linguistic microevolution') by testing a population genetic clustering method for analyzing the 'population structure' of Finnish dialects. We compare the results with traditional dialect divisions established in the literature, and with K-medoids clustering, which is free from biological assumptions. The results are encouragingly similar with each other and agree with traditional views, suggesting that population genetic tools could be a useful addition to the dialectological toolkit. We also show how the results of the model-based clustering could serve as a basis for further study.

Keywords: microevolution, quantitative dialectology, population genetics, Structure software, K-medoids, Finnish dialects

1 Introduction

With written accounts dating back to at least the 14th century (Heeringa 2004), dialects have generated a great deal of interest among language researchers over the years, with systematic dialect study (dialectology) stemming back to the late nineteenth century (Chambers & Trudgill 1998). Dialectology has generally focused mainly on traditional (non-quantitative) research, although statistical analyses have also gained foothold from the 1950s onwards (Chambers & Trudgill 1998). Current dialectology includes a number of computational approaches, including multivariate analyses, dialectometry, Levenshtein distances, clustering and multidimensional scaling (e.g. Heeringa 2004; Leino *et al.* 2006; Hyvönen *et al.* 2007; Leino & Hyvönen 2008), and new methods continue to be developed.

Quantitative methodology from biology might provide a useful addition for the repertoire of dialectological tools. Biological methods have gradually seeped into other linguistic fields, most notably historical linguistics, where they are used to study 'linguistic macroevolution' – e.g. language classification, divergence history, and the forces driving linguistic divergence (e.g. Gray & Atkinson 2003; Lee & Hasegawa 2011; Honkola *et al.* 2013; Syrjänen *et al.* 2013; Lehtinen *et al.* 2014). In studying linguistic macroevolution languages are regarded as roughly analogous to species. It might be possible to take this analogy a step further, as in a similar way as languages have language-internal variation which may be clustered to dialects, species also have internal variation which may be clustered to populations. Studies on biological populations focus specifically on studying this variation with specific tools in their own research fields (including e.g. population genetics and population ecology). This provides an interesting possibility of approaching dialects from a 'microevolutionary' perspective by adopting approaches from disciplines designed to explore within-species variation to the study of intra-lingual varieties. In this paper we examine the applicability of this 'microevolutionary' approach to dialect studies.

Our data comes from an atlas of Finnish dialects collected at the beginning of the 20th century (Kettunen 1940a), which we analyze using population genetic clustering, and compare it with a generic distance-based clustering method. Both are also compared against dialectological studies. The study focuses largely on methodological exploration; Finnish dialects, with their extensive study history, provide a good baseline for this. While clustering plays a large role in this study, our main purpose is not to determine the number of dialects best supported by population genetic analyses but rather to achieve an in-depth view of dialect clustering as the initial stage for more advanced analyses.

We begin by briefly introducing the underlying theoretical framework – i.e. languages and dialects from an evolutionary perspective. Following this, we take a look at earlier dialectological research, outlining how Finnish has been divided dialect-wise in the past. We also take a look at quantitative dialectology related to Finnish dialects. Following this we introduce the data and the methods employed in this paper. Finally, we evaluate and discuss the results in the light of Finnish dialectology and conclude by presenting some examples of analyzing dialects with population genetic tools that go beyond the clustering step.

2 Background

2.1. Linguistic evolution: how languages and dialects resemble species and populations

The present study does not require extensive familiarity with the full range of analogies and similarities proposed between biological species and languages, which date back all the way to Darwin. Here, we focus on what we consider to be the three most important analogies or similarities, which serve as the core for applying biological methodology to study language data. Firstly, both have discrete heritable units. Secondly, these heritable units are packed in spatiotemporal "containers" – the individuals, which are typically structured in groups (populations). Thirdly, the individuals and the populations are susceptible to internal and external forces that affect the frequencies of the shared heritable units over time. In the following we discuss these points more in depth. A more in-depth evolutionary analysis of languages, largely compatible with what is discussed here, can be found in Croft (2000); a concise selection of analogies that is also generally compatible with the model we describe here can be found in Pagel (2009).

The heritable units in biology – the genetic information carried by organisms, e.g. genes, alleles, nucleotides and amino acids – transfer primarily vertically from parents to offspring through genetic inheritance. Additionally, horizontal gene transfer has been occasionally found to occur (e.g. Gasmi *et al.* 2015) and actually the early stages of life were presumably characterized by extensive horizontal gene transfer among prokaryotes (e.g. Campbell *et al.* 2008). The heritable units in languages – e.g. words, phrases, constructions – are transmitted via communication between individuals. This shows a significant difference between biological species and languages – that is, in biology the heritable units are carried over to a newly created individual (the offspring), while their linguistic counterparts are carried over to an existing speaker. What is similar in both cases is that the heritable units are continuously transferred between individuals or organisms that are part of that system, making it possible for the heritable units to persist to a considerable degree across generations. A characteristic that both these systems share, and one which makes them 'evolutionary', is that the heritable units and also their frequencies of occurrence change in time.

Both languages and species involve individuals serving as carriers for the heritable units. In the case of sexual reproduction and linguistic transmission the transfer of heritable units necessitates interaction between the individuals. In both processes the individuals do not interact uniformly with all the other individuals, so the heritable units within a single species or language are distributed unevenly. For this reason it is possible with both sexually reproducing species and languages to identify subgroups of individuals whose heritable units (genetic material or linguistic information) is closer to each other than it is to that of the other individuals. In biology these are groups are called 'populations', and they can be regarded as being analogous to e.g. dialects. Individuals in a biological population are generally capable of interbreeding with the members of another population, but tend to interbreed more within their own population. Over time this forms a clear detectable pattern in the shared heritable units. This is largely analogous with how the speakers of a dialect are more likely to communicate with each other although they are generally capable of communicating with the speakers of other dialects. Due to this a distinct pattern of linguistic units – dialect, or more generally a linguistic variety – emerges. [1]

The third essential similarity between languages and biological species, which is also true of within-species populations and intra-lingual varieties, is that the differences that we can observe in the heritable units can essentially be modeled as a combination of unpredicted (non-directional) changes and directional changes (selective pressures) (Croft 2000; Levinson & Gray 2012). The selective pressures are of course different for the two; for instance, social factors arguably act as an important type of selective pressure in the linguistic realm. Their closest counterparts in the biological realm could be within-species interactions, such as competition, which however do not have as much prominence as social selection has for languages. Although the biological and the linguistic realm generally operate under different rules and are influenced by separate selective pressures, they are not entirely disconnected; the speakers themselves are entities in the biological realm, and therefore also subject to biological pressures. However, we must also remember that humans counteract many biological selective pressures with cultural adaptations, making the overall picture of different selective pressures quite complex.

The aforementioned similarities between languages and intra-lingual varieties and biological species and within-species populations serve as the basis with which languages can be modeled under an evolutionary linguistic framework. The study of the evolutionary processes involving species (phylogenetics) and within-species populations (population genetics) are two distinct sub-disciplines of biology that share a general theory, but use different approaches – one designed to reveal a tree or network describing a large-scale pattern of accumulated changes and the other to describe minute differences between individuals of the same species. In a similar way we describe the present study, which focuses on modeling dialects with population genetic tools, as the study of "linguistic microevolution", to contrast it with studies focusing on differences between languages – "linguistic macroevolution" – such as phylogenetic linguistics.

2.2 Finnish dialect division

Subjective accounts of Finnish dialects are as old as written Finnish, with one of the earliest descriptions found in Mikael Agricola's foreword for the New Testament (Agricola 1548). Systematic dialect research is generally considered to have begun around the nineteenth century, motivated partially by growing interest in national history and fieldwork focusing on collecting oral tradition (Hovdhaugen *et al.* 2000). Dialectology remained among the most active topics in Finnish linguistics until the mid-twentieth century, when variationist studies shifted more towards sociolinguistics (Hurtta 1999). As a whole, the bulk of Finnish dialectology is traditional work, the large majority being detailed descriptions of individual dialects and dialect areas based on fieldwork, although some works have focused on Finnish dialect variation as a whole, such as Kettunen (1930, 1940a, 1940b), Hakulinen (1950), Rapola (1969) and Hormia (1978).

[1] The similarity between language-internal varieties and biological populations can be seen in Croft's (2000) analogy between traditional geographical dialects and geographical races, as well as his analogy between social networks and biological demes. Similarly, Pagel (2009) likens dialects and dialect chains with geographical clines.

There is a fairly good consensus on categorizing Finnish dialects. The most common general division splits the language into two principal dialect areas, Eastern and Western. This dichotomy was characterized already in the eighteenth century by Vhaël (1733), and became the default division in early nineteenth century (Rapola 1969, Wiik 2004). It is regarded as the clearest general division of Finnish dialects, and also serves as the foundation for more fine-grained divisions, particularly those that emphasize morphological and phonological features.

The eastern and western dialects are often subdivided into seven or nowadays often eight main dialects (e.g. Itkonen 1964, Savijärvi & Yli-Luukko 1994), which are generally clear, although slight variation can be found (e.g. Mielikäinen 1991, Leskinen 1992). Itkonen (1964; 1989) is often considered the 'gold standard' of the eight-way divisions, splitting the western dialect area into Southwest, Southwest transitional, Häme, South Ostrobothnia, Middle / North Ostrobothnia and Far North, and the eastern dialect area into Savo and Southeast (Fig. 1).

(Fig. 1 here)

Although the two-way division stands as the default division for Finnish, three-way divisions have also been suggested. Based on Rapola (1969), one of the oldest of these is from 1777, when Erik Lencqvist suggested a division of Finnish into 1) Turku dialect, covering parts of the Southwest and Southwest transitional dialects, 2) Ostrobothnian dialect, which also included Häme and 3) Savo dialect. In essence, this suggested Itkonen's (1964) Southwest as a main dialect rather than a subdivision. The three-way division has generated some later discussion by Mielikäinen (1991) and Paunonen (1991; 2006), who have suggested that e.g. synchronic typological features support this kind of division. Another kind of three-way division splits Finnish into eastern, western and northern areas, with the northern area being essentially a mixture of eastern and western influence. This division was originally proposed by Warelius (1848), and has been discussed later in e.g. Leino *et al.* (2006) and Hyvönen *et al.* (2007). The east-west-north trichotomy has been suggested to be more prominent at the lexical level, whereas the two-way division (east-west) is more prominent at morphological and phonological levels.

There are also some grounds for suggesting four principal dialect areas. Paunonen (2006), going beyond Lencqvist's trichotomy, suggests that from a synchronic standpoint Finnish should be divided into 1) Southwest dialects, 2) Western dialects (covering Southwest transitional dialects, Häme dialects and South Ostrobothnian dialects), 3) Eastern dialects (covering Savo and Southeast dialects), and 4) Northern dialects, covering Middle / North Ostrobothnia and Far North.

2.3 Quantitative dialect studies of Finnish

In this section we look at four quantitative, ('dialectometrical') works on Finnish dialects. The works we discuss here include Wiik (2004), Leino *et al.* (2006), Hyvönen *et al.* (2007) and Leino & Hyvönen (2008), all of which – like the present paper – explore Finnish dialects as a whole.[2] These do not represent the whole range of methods within

[2] Alongside these studies we should also mention Embleton & Wheeler (1997; 2000), who have contributed to quantitative dialect studies of Finnish by creating the digitized version of the Dialect Atlas

quantitative dialectology; more on the subject can be found in e.g. Chambers & Trudgill (1998), Palander (1999), Nerbonne & Kretzschmar (2003) and Heeringa (2004).

Wiik (2004) is perhaps the most comprehensive quantitative take on Finnish dialects so far. He presents a numerical interpretation of the Dialect Atlas of Finnish (Kettunen 1940a), executed by counting co-occurring isoglosses by drawing each of the dialect atlas maps on transparent slides, visually inspecting the stacked slides and compiling progressively larger composite slides until he ended up with a summary of the entire atlas. Wiik reflects his calculations mainly against standard dialect divisions: the east-west dichotomy and Itkonen's (1964) eight-way division (see section 2.2). The subgroups of each dialect area are also discussed carefully, and each of these is reflected against Wiik's measurements. The work also outlines 'core areas' for each dialect, based on the coverage of the dialect features that have been considered 'primary' for each dialect. In general, the work does not attempt to redefine the dialect division from a quantitative perspective, but rather explore and refine the eight-way dialect division. To some extent the work resembles Séguy's dialectometrical additions to the *Atlas Linguistique de la Gascogne* (see e.g. Chambers & Trudgill 1998 for an overview). What makes the work quite impressive is that it has been done mostly manually, using the paper version the dialect atlas.

The main focus of Leino *et al.* (2006) and Hyvönen *et al.* (2007) is on lexical variation, making it an interesting exception among the predominantly morphological and phonological dialect studies of Finnish. They employ multivariate analyses adopted from data mining, including principal component analysis, independent component analysis, multidimensional scaling and distance-based clustering, to explore the distribution maps produced in the course of editing the Dictionary of Finnish Dialects (Tuomi 1989). Their results agree surprisingly well with traditional dialect studies, with the exception that the lexical data appears to be better generalized using a north-east-west trichotomy similar to the one suggested by Paunonen (1991) (see section 2.2), and not the east-west dichotomy.

Leino & Hyvönen (2008) expands the work started in Leino *et al.* (2006) and Hyvönen *et al.* (2007) to also cover morphophonological variation, using data from the digitized Dialect Atlas of Finnish (Embleton & Wheeler 1997; 2000) alongside the Dictionary of Finnish Dialects. Like in their previous works, they explore various approaches for analyzing the data: factor analysis, non-negative matrix factorization, aspect Bernoulli, independent component analysis and principal component analysis. Based on earlier work they prefer these methods over distance-based clustering, such as K-medoids, because they do not impose sharp boundaries and are thus a more natural choice for dialects. The work highlights how differently the methods work with these two datasets, which differ significantly from one another with respect to both content and quality. Based on their tests they present further results using factor analysis, which they found to perform reasonably well with both datasets. These results, perhaps more than anything else, highlight how lexical and morphophonological data reveal different but not entirely conflicting variation patterns.

Reflecting how the aforementioned works relate to this study, the data we examine is the same that Wiik (2004) and Leino & Hyvönen (2008) used, although differently represented, and our analyses represent partitional (non-hierarchal)

of Finnish, and who have used it to explore MDS techniques for visualizing dialect information. Another noteworthy dialectometrical investigation which we did not include here is the study by Palander *et al.* (2003), focusing on the regional dialects of Savonlinna.

clustering, also employed as part of Leino *et al.* (2006) and Hyvönen *et al.* (2007). Notable differences to the existing studies include our almost exclusive focus on partitional clustering, and our usage of population genetic thinking and tools.

3. Materials and methods

3.1 Finnish Dialect Atlas

The data used in the analyses comes from the Dialect Atlas of Finnish (Kettunen 1940a), compiled by Lauri Kettunen[3] in the 1920s-1930s. During this period he travelled across Finland interviewing informants and documenting local regional speech. Based on Kettunen's travelling memoires (Kettunen 1960) he generally interviewed at least two informants per municipality, and made efforts to find more in ambiguous cases. He looked for informants by consulting local priests and visiting old people's homes and prisons, searching for old and uneducated locals that had been living in the area for their whole life. The resulting dialect atlas (Kettunen 1940a) mainly documents the distribution of morphological and phonological phenomena, with less information about lexical variation. It was accompanied by an explanatory book (Kettunen 1940b), and is closely related to his earlier dialect book (Kettunen 1930), which was intended to serve as an introduction to the atlas.

The atlas covers 213 linguistic features, presented as separate maps (Fig. 2), with information from 525 sites (municipalities), covering all of Finland except exclusively Swedish-speaking areas, located in the western and southern coast of Finland. It also covers Finnish-speaking areas in Ingria (Russia), Norway and Sweden, as well as Karelian-speaking areas in pre-WWII Finland. Each map shows the municipality-wise distribution of the different variants of linguistic features. The atlas does not document responses from each informant; the data points represent the combined information from all the informants from of that municipality. The number of variants per page range from 2 to 15, and the number of variants per municipality ranges from 1 to 4. Embleton & Wheeler (1997) estimated that the atlas covers up to 36 times as many dialect "facts" as the Survey of English Dialects.

The basic study unit of the atlas is "the dialect variant in a municipality". While theoretically it would cover 111,825 study units, this is not the case; data is missing from 8.1 % of the study units. Especially certain peripheral areas have gaps in the data; for instance, there are altogether twelve municipalities with less than 100 dialect features. These include six municipalities in Northern Lapland, three mainly Swedish-speaking municipalities in the coast of Ostrobothnia, two islands in the Baltic Sea, and a municipality in Karelia. The area with most gaps appears to be Lapland. Area-wise this is significant, as the municipalities in that area are fairly large. Despite the gaps, we analyzed the data as a whole in this study, using data from all the map pages and municipalities.[4]

Our analyses required a computerized version of the atlas, available thanks to the work by Sheila Embleton and Eric Wheeler as part of the Finnish Dialect Atlas Project, funded by Social Sciences and Humanities Research Council of Canada in co-operation

[3] Information on South Ostrobothnia was not collected solely by Kettunen but instead taken from Laurosela's work on the South Ostrobothnian dialect (1922).

[4] This decision is not without problems. In addition to the gaps within the data, Wiik (2004) has pointed out that the features in the atlas range from very generic to very specific, and including them as such gives both types equal weight. Some complex phenomena are also meticulously documented in the atlas, spanning multiple map pages. Some of the recorded features in the atlas also concern only a small area of Finland, meaning that the data for these features is intentionally missing from most municipalities.

with the Institute for the Languages of Finland (Kotus) (Embleton & Wheeler 1997; 2000). An additional round of error checking for the digitized atlas was done by one of the authors of this paper. An online version of the data was published by Kotus in 2015 (http://avaa.tdata.fi/web/kotus/aineistot).

(Fig. 2 here)

3.2 Data formatting

For the analyses the dialect data needed to be in a format compatible with our two analyses – Structure and K-medoids (see section 3.3). Here we explain how this was accomplished.

Genetic data - more specifically, alleles (variants of a gene) sampled at specific loci (location of the gene) from several individuals of the same species serve as Structure's input data. In order to infer dialect populations with Structure we treat the 525 municipalities as individuals, the 213 map pages (each of which describes the distribution of the variants of a particular dialect feature) as genetic loci and the variants within each map page as alleles.

Biological organisms differ in how many alleles per locus their genotype includes. Mammals are generally *diploid*, i.e. for each locus they have two alleles: one inherited from the mother and one from the father. If a diploid organism has inherited the same allele for a certain locus from both parents, it is *homozygous* for that locus; if it has two different alleles, it is *heterozygous*. There also exist organisms (e.g. male bees, wasps, ants, certain life stages of algae, ferns and mushrooms) that are *haploid*; these only have one allele in each locus. There are also *polyploid* organisms, with 3 or more alleles.

Most study units (94.3 %) in the dialect atlas are "haploid", with only one variant of a linguistic feature per municipality. For example, on map 8 of the atlas (Fig. 2), representing variants for "forest", the easternmost municipalities are marked with just one symbol (horizontal curvy lines); in the digitized data this is represented by 3 (the third box in the legend). In contrast, some of the municipalities in the east are "diploid", as they are marked with both red triangles and crosses. In the digital version this is marked as (12, 13, i.e. the twelfth and fifteenth boxes in the legend). In total 5.6 % of the study units are "diploid". A small number of the study units (0.1 %) included 3 or 4 overlapping variants. Of these the third and fourth variant were excluded for simplicity's sake.

In this work we turned the data into two forms, haploid and diploid, and analyzed both of them. Our main focus was on the diploid representation, as it covers almost all of the variation in the atlas. In contrast, the haploid version only covers the first marked variant for each linguistic feature. Following Structure's guidelines, for the diploid coding the study units with two variants were left as they were (e.g. 12, 13); in cases with only one variant the variant was duplicated – e.g. 2 became (2, 2).

3.3 Clustering methods

3.3.1 Model and distance based clustering

Placing data into meaningful subgroups has been of great interest in many fields, including dialect studies and biology, resulting in a wide selection of clustering methods (e.g. Kriegel *et al.* 2009) based on different principles. Our focus in this study is on partitional clustering approaches, which produce non-hierarchal groups and are often used in inferring population structure from genetic data. Clustering can be roughly divided into two types: model-based and distance-based (Pritchard *et al.* 2000), both of which are used in this paper, and introduced in the following section.

In model-based methods, each cluster is assumed to be generated by a specific probability model. Model-based clustering aims to infer the probability models representing the clusters from the data itself, and place the data as best as possible into these clusters. Model-based methods tend to be computationally intensive, and have only recently gained foothold in research through tools based on Bayesian MCMC methods. In this study we use a model-based clustering tool called Structure (Pritchard *et al.* 2000), designed to infer population structures from genetic data. The method has been applied earlier to cluster languages and language varieties (Dunn *et al.* 2008; Reesink *et al.* 2009; Bowern 2012); here we use it specifically to study intra-lingual variation.

In contrast with model-based clustering, distance-based clustering is more straightforward; a distance or a similarity function is specified and used to measure distances between data points and cluster together points close to each other. These approaches are older than model-based methods, and are generally computationally faster. The distance-based method we use here is K-medoids (Kaufman & Rousseeuw 1987), a method which has also been applied for the study of Finnish dialects earlier (Leino *et al.* 2006; Hyvönen *et al.* 2007).

3.3.2 Structure

As a population genetic clustering approach we use Structure (Pritchard *et al.* 2000), a model-based software that uses Bayesian methods to infer biological populations from genetic data (see Beaumont & Rannala 2004 for a general overview). Structure is not the only software of its kind; tools built on similar principles include e.g. BAPS (Corander *et al.* 2003) and TESS (Chen *et al.* 2007).

As is explained in Pritchard *et al.* (2000), Structure is designed to analyze a set of alleles (variants of a gene) sampled from individuals of the same species. The individuals can be assumed to originate from one ancestral population (a term that is essentially interchangeable with the term ‘cluster’), or have an admixed origin from several populations. Structure treats the ancestral populations and the placement of the individuals into the populations as separate unknown parameters that it endeavors to estimate simultaneously. The ancestral populations are represented by a model that specifies the allele frequencies for each locus, i.e. how widespread each allele is within each population. Structure infers a division by assigning a population for each of the data points, and then estimates the overall likelihood of the solution using the allele frequencies it has inferred for the ancestral populations. Then, following standard Bayesian MCMC methods, one of the unknown parameters is modified while the remaining parameters are retained, and a likelihood score for this new solution is estimated. If the new solution has a higher likelihood than the previous solution, it is accepted; if not, it is accepted with a probability of A/B, where A is the estimated likelihood of the current solution and B the estimated likelihood of the previous solution. The algorithm repeats the procedure, randomly modifying another unknown parameter and calculating the likelihood of the new solution and comparing it to that of the previous solution, and so on. Continuously repeating the aforementioned steps and storing the results at predefined intervals yields a distribution of solutions, and gradually converges on the most optimal solution or solutions. Each finished analysis includes a likelihood estimate of the data when divided to K populations, which Structure summarizes from the entire MCMC run. Since the analysis generally starts with many unknowns for which arbitrary starting values are chosen, the iterations at the beginning of the analysis are not informative and possibly even misleading. For this reason, these initial results, referred to as 'burn-in', are discarded (Pritchard *et al.* 2000).

With the admixture model Structure produces *soft* or *fuzzy* populations by assigning each data point (municipality) a degree of membership (IC value) in each ancestral population. This makes it possible to infer a mixed origin for the data points. In this study we use this model, as it is naturally suited for dialects, which may often involve gradual transitions from one variety to the next. In contrast, the distance-based K-medoids clustering only infers *hard* clusters, where a data point can only belong to one of the clusters.

Structure requires the user to specify how many populations to infer, so it is useful to be able to determine how many clusters best explain the data. We will discuss methods for determining the optimal number of clusters in section 3.4.

The analyses for this study were run for all K (the number of populations) from 1 to 20, with each analysis repeated 20 times to ensure the consistency of the results. The burn-in period was set to 10,000 generations and the number of MCMC repetitions after burn-in was set to 100,000 generations. The Admixture model was used, allowing individuals to originate from more than one population. As was already mentioned, we prepared two representations of the dialect data, diploid and haploid, and analyzed both, largely to see if this change in the nature and amount of the variation data affected Structure's results. In the results section we focus primarily on the diploid results, which has better overall coverage of the linguistic variation. The haploid results are not given in the main text, but their comparisons with the diploid results are summarized in the appendices.

The results of the diploid Structure analyses are presented in two ways. Firstly, for each K, the repetition with the highest likelihood score is visualized on a map. Secondly, the repetitions of each K value, excluding clear outliers, were summarized using Structure Harvester (Earl & vonHoldt 2012) and CLUMPP (Jakobsson & Rosenberg 2007; see section 4.3.1).

3.3.3. On the biological assumptions of Structure

There are two notable biological assumptions embedded in Structure's algorithm that deserve discussion. Firstly, Structure infers populations that correspond with the 'Hardy-Weinberg equilibrium' (Pritchard *et al.* 2000), or HWE, as well as possible. This is an idealized state; in order for a biological population to be in HWE, it would need to be for example infinite in size, unaffected by any kind of natural selection and reproducing completely randomly (Hamilton 2009), a state which is not a valid generalization of real life populations on a longer time span.

It would also be unrealistic for language variants to remain in HWE, as it would require for example a random spread of linguistic variants across speaker populations and that the frequencies of linguistic variants should not be affected by any 'selective' force such as social selection. These requirements are not, however, often met as for example linguistic variants or innovations generally have a certain geographical pattern, as language speakers in geographical proximity often communicate more. Additionally, languages are constantly changing for example due to contact-induced changes and innovations, which are not necessarily random. Therefore, the longer period of time we are observing, the less plausible it is to assume that languages or dialects would have remained the same and retained a HWE state.

For Structure this HWE criterion reflects above all the fact that Structure's model does not cover mutation (or innovation), so the populations it infers are such that would be the result of a set of already existing alleles mixing at different ratios. From the perspective of languages this could be thought of as a situation where variation comes about predominantly through a process like *intraference* (Croft 2000), where already existing linguistic features are adopted by speakers of different dialects at different ratios.

Populations where the allele frequencies have remained unchanged (i.e. populations that are in HWE) would in essence represent ancestral populations, i.e. populations representative of the linguistic situation spanning far back in time. Correspondingly, if the data is not in HWE, as most likely is the case with the language data, these interpretations cannot be made and we need to assume that they reflect a population division that is, on a temporal scale, fairly close to the age of the data itself. With this in mind, the HWE assumption does not limit what we can analyze; however, if HWE is unlikely, we need to avoid making strong interpretations of the results that would necessitate it, such as assuming an unrealistic time depth.

Another assumption that Structure makes is that the variables in the data are independent: loci should be in 'linkage equilibrium'. With genetic data when certain gene combinations occur together more often than they would randomly, they are said to be in 'linkage disequilibrium'. This state may arise through several mechanisms, such as physical linkage, where loci are situated in the same chromosome and close to each other and thus the alleles in these loci tend to be inherited together. If the loci are further away from each other the alleles are more likely inherited independently and thus, also more likely to be in linkage equilibrium (Hamilton 2009). Linguistic information does not resemble genetic information in that features would be stored physically close. On a cross-linguistic level, implicational universals, i.e. features that frequently co-occur across languages, could be seen as analogous to linked loci. However, Kettunen's dialect data covers features that highlight differences between Finnish dialects, which are too specific to be universal features. Thus, here linkage would essentially be the presence of systematically correlated features of Finnish within the atlas. Indeed, we could expect certain characteristics within the atlas to have a degree of linkage, such as the meticulously documented instances of consonant gradation, which have been suggested to carry redundance by other studies of the dialect atlas, such as Wiik (2004).

Similarly as we did not exclude any data points from the analyses based on uneven coverage, we also refrained from excluding map pages based on assumed linkage. For language data there are no attested methods exist to study linkage from our type of data.[5] However, we created one kind of ad hoc test with which to measure the extent of linkedness between the map pages in the atlas. The method which goes

[5] Some biological linkage tests exist, in particular Lewontin's D and its derivatives, based essentially on how much the allele combinations from two loci diverge from the expected frequencies of randomly combined alleles. We tested the D' (Lewontin's normalized D) metric for language data but found out that this metric is not directly applicable for language data.

through all the data point (municipality) pairs (x, y) on each pair of map pages (a, b), and checks if the municipalities x and y are linguistically identical (i.e. the same set of dialectal features are used in both x and y) on map page a, and does the same for map page b. The cases where x and y are marked identically on at least one of the pages are counted as being “potentially linked” (Lp), and the cases where x and y are linguistically identical on page a as well as page b are counted as being “actually linked” (La). The calculation discards any cases where x or y have no marked features on either map page. After we have checked the linguistic features for all the possible municipality pairs on given pair of map pages and recorded La and Lp from them, we estimate the amount of linkage on that map page pair as La/Lp, i.e. the number of “actual linkage” cases divided by the number of “potential linkage” cases. Thus, the metric essentially calculates how many pairs of municipalities that had the potential of being identically marked on the two pages under inspection (by being marked with identical dialect features on either page) were actually marked identically on both pages.

All of the map page pairs were compared in this way, with the help of a custom-made Python script. The results were visualized with R (R Core Team 2014) using *heatmap.2* from the *gplots* package (Warnes *et al.* 2014). It should be noted that the linkage estimation test remains fairly rough, with considerable room for improvement.

3.3.4 K-medoids

K-medoids (Kaufman & Rousseeuw 1987) is a distance-based clustering method that, like Structure, creates non-hierarchal groups. It is essentially an improved type of K-means clustering, being less sensitive to outliers than its predecessor. K-medoids has been used to explore lexical data from the Dictionary of Finnish Dialects (Leino *et al.* 2006; Hyvönen *et al.* 2007).

As its input K-medoids takes data points represented as a set of features in numerical form (or in mathematical terms, data points represented as feature vectors in n-dimensional space, n being the number of features). K data points are randomly selected as *medoids* (centers for the groups), and the distance between each medoid and data point in the data set is calculated, and the points are assigned to the groups closest to them. After the points are assigned to groups, the algorithm calculates the total distance from each point to all the other points in the group. If this distance is lower than the combined distance from the original medoid point to the other points in the group, the point with the lowest combined distance becomes the new medoid. If the medoids changed, the algorithm re-evaluates each data point against the new medoids, and reassigns them to new groups as necessary. The re-evaluation of the medoid points and the reassignment of the data points continues until the groups do not change any more, or until the algorithm has gone through a predefined number of iterations.

The analyses were done using the same data that was used for Structure’s diploid analyses, although it had to be represented differently. K-medoids cannot account for missing data points, so for K-medoids missing (empty) characters and absent linguistic features were marked identically as 0s (Structure’s data representation, on the other hand, retains the distinctions between missing and absent features). The R package *cluster* (Maechler *et al.* 2014) and its command *pam* was used for the analyses, using the default settings. Like Structure, K-medoids also requires the user to specify the value of K. We asked K-medoids to divide the data into 2-20 clusters. Repetitions of the analyses suggested that the K-medoids clusterings were consistent.

3.4 Estimating optimal K values

Both K-medoids and Structure rely on the user to specify the number of clusters or populations. Because of this, it is important to be able to estimate which partitioning best explains the data.

The optimal K value in Structure can be estimated more formally using the ΔK metric and less formally from the mean log likelihood (Evanno *et al*. 2005). The mean log likelihood is calculated (after excluding outliers) by averaging the log likelihood values from all the repetitions for each K while ΔK displays how much the mean log likelihoods change on each K value when compared to the neighboring K values (K-1 and K+1). Thus, when the mean log likelihoods differ dramatically in a given K value compared to the neighboring K values, ΔK is high. Mean log likelihoods and ΔK were calculated for each K with the R package *pophelper* (Francis 2014).

These two metrics may be used jointly to estimate the kind of partitioning that would best explain the data. Commonly the mean log likelihood values are small with small K values, and more or less plateau for larger K. (Pritchard *et al.* 2010). In this kind of situation it is suggested that the smallest K value with which the K values plateau is usually the one explaining the data the best. This point should also be the one that is supported in the ΔK calculations as the difference between neighboring values is supposed to be highest when reaching the plateau.

To estimate different K values for K-medoids we used the silhouette method (Rousseeuw 1986). It involves examining the relationship of within-dissimilarity (the average distance among the data points in the cluster) and between-dissimilarities (a data point's average distance to points in a different cluster) of the data points. A silhouette value is expressed as a ratio of within-dissimilarity to the lowest between-dissimilarity, so essentially it describes how well a data point fits its current cluster compared to the neighboring cluster. The silhouette value varies between -1 (a poorly classified point, i.e. much closer to the neighboring cluster) and 1 (a well-classified point, with considerable distance to the next best cluster). Across the entire result we can examine *average silhouette width*, the average of all the silhouette values for a clustering, which shows how well the current K value generally describes the data. The R package *cluster* (Maechler *et al.* 2014) was used to calculate silhouette values.

3.5 Visualization

Structure's results for each municipality are given as a set of membership coefficients (IC values, see Fig 3); each municipality gets an IC value for each inferred population. The IC values can be regarded as percentages that sum up to 100%, and show how the inferred populations are mixed on each data point (municipality). E.g. when we infer three populations, a municipality could have a 70 % membership for population A, 20 % membership for population B and 10 % membership to population C. These membership coefficients allow flexible visualization.

The standard way of visualizing results such as these would be to use Structure's bar plot visualization (Fig. 3a), which can show the full mixture of the dialectal characteristics for each municipality. However, even though this type of visualization is very detailed, it is not very illustrative if you are interested in the geographical location of the studied units. To obtain visual clarity for our main results we chose to group the membership coefficients into two distinctly colored main classes, which were plotted onto a map. These two classes were core dialects (municipalities with IC > 0.75; colored with saturated colors) and transitional dialects (municipalities with IC = 0.50 - 0.75; colored with less saturated colors); this is shown on Fig. 3b. Notably, this could also have been done with more than two classes (Fig 3c), but we felt no need for it for the purposes of this study.

These colored clusters were plotted on a base map representing Finnish municipal boundaries in the 1920s, digitized with modern Finnish national topographic database elements using the geographic information system ArcGIS. The base map was prepared by Ilpo Tammi for the BEDLAN project. The digitization was mainly based on the facsimile of *Suomen kartta 1920* (Harju 2009) and the *Atlas of Finland 1925* (Geographical Society of Finland 1928). Supplementary sources, chiefly the Atlases of Finnish ethnic culture (Vuorela 1976, Sarmela 1994), were used to identify the historical boundaries for extraterritorial areas linked with Finnish dialects.

Our visualization of the results (Figs 7-9) has the setback of displaying just the highest membership coefficient values instead of the entire mixture. Visualization showing the full mixture of the IC values is also possible, but this type of visualization would be fairly difficult to interpret for a geographical area of the size that we are studying. Fig. 4 gives examples of this type of visualization.

(Fig. 3 here)
(Fig. 4 here)

Unlike Structure, K-medoids places each data point (municipality) explicitly into one cluster, and consequently does not produce membership coefficients which would show mixture proportions. These can be visualized on a map by simply giving a unique color for each cluster. In the visualizations the K-medoids clusters match with the corresponding populations in Structure as well as possible, and use the same saturated colors that are used for the “core dialect” class (IC > 0.75) in Structure. As K-medoids does not membership coefficients, its clusters are more clear-cut than the populations produced by Structure.

4. Results

4.1. The optimal number of clusters

The average likelihoods from the Structure analyses (Fig. 5a) increase gradually as the K value increases from 2 to 14 without reaching a clear plateau. With values exceeding K=14 the likelihoods begin to fluctuate across runs and their mean values decrease. The changes in the likelihood values are quite small – this essentially means that the K values between 2 and 14 explain the data almost equally well, with K=14 being the best by a small margin. Although likelihood values do not emphasize any certain K value well above all others, the *ΔK* values (Evanno *et al.* 2005) (Fig. 5b) show a notable peak with K=2, suggesting that a division into these two clusters would represent the best uppermost division of the data.

(Fig. 5 here)

The average silhouette widths, calculated for the K-medoids clusterings, range from 0.15 to 0.24 (Fig. 6), meaning that the clusterings are neither too good nor exceptionally bad. This is generally in line with Structure's likelihood values – that is, no K value explains the dialect data exceptionally better than the others. Unlike Structure's log likelihoods, the silhouette values remain fairly stable also beyond K=14. K=3 appears to be the least optimal of the lower K values. The average silhouette widths stabilize at K=6 and beyond. K=16 has the highest average silhouette width, 0.24, albeit by a very small margin.

(Fig. 6 here)

Except for *ΔK* peaking with K=2, the support metrics do not clearly favor any specific K value. The high silhouette values above K=14 suggests that exploring clusterings at higher numbers may be of interest, whereas Structure's likelihoods suggest that they are of less interest. Here we decided to focus on the clusterings within Structure's high likelihood area (K=2-14).

4.2 Dialect clusters

The populations inferred by Structure were generally in line with the clusters found by K-medoids. Divisions with K values 2–8 (Fig. 7) were most similar between the two. Divisions with K=9–14 (Fig. 8) showed more variation across analyses, in especially in the order in which the clusters or populations appeared as K increased.

In what follows we will examine the dialect divisions in detail, beginning with the more stable divisions (K=2–K=8), followed by the less stable ones (K=9–K=14). We also compare these with CLUMPP visualizations, which align repetitions of Structure runs as well as possible, revealing solutions that disagree across repetitions. In general, the clusterings were clear, except for some northeastern municipalities that appeared as transitional areas for random dialect clusters from K=6 onwards in Structure, and from K=3 onwards in K-medoids. This was likely caused by the scarcity of linguistic features in those municipalities, as they seemed to cluster together with less documented municipalities with higher K values in K-medoids (see section 4.2.2 for more details).

4.2.1. Divisions from K=2 to K=8

Except for K=3, Structure and K-medoids suggest essentially identical divisions with all but one K value between 2 and 8 (Fig. 7). The first division, K=2, is between the Eastern and Western dialect groups. With K=3 the Eastern group remains unchanged in both, but Structure separates the Southwest dialect area (red in Fig. 7) from the Western dialects while K-medoids splits the Western dialects to Middle / North Ostrobothnia + Far North (blue) and the rest (purple). In K=4 the Eastern dialect group remains intact while in the Western dialect area Häme (purple) appears next to Southwest (red), and the northernmost cluster (blue) now roughly covers Ostrobothnia and Far North.

With K=5 the Eastern dialect area splits into two clusters: Southeast (dark gray) and Savo (green). Increasing the K value to six separates South Ostrobothnia (orange) from Middle / North Ostrobothnia + Far North (blue). K=7 separates Southeast Häme + Päijät-Häme (brown) from the main Häme dialect, while K=8 makes the Far North dialects (olive green) a separate cluster. To sum up, a division to 8 clusters gives two Eastern dialects (Savo and Southeast) and six Western dialects (Southwest, Häme, Southeast Häme + Päijät-Häme, South Ostrobothnia, Middle / North Ostrobothnia + North Kainuu + Kemijoki, and Far North) in both analyses.

(Fig. 7 here)

4.2.2. Divisions from K=9 to K=14

Most of the new clusters between K=9 and K=14 (Fig. 8) are subdivisions of Eastern dialects, with less within the Western dialect area. The analyses also start to disagree more with higher K values.

Some clusters appearing with these K values are fairly stable, including the Southwest transitional dialect area (light green) – present with K values 10-14 with both Structure and K-medoids – and Päijät-Häme (bright blue) – appearing with K=11-14 in Structure and with K=14 in K-medoids. With its appearance, the brown cluster, formerly covering Southeast Häme and Päijät-Häme, decreases in size to cover just Southeast Häme.

Among the less stable clusters we find Central Karelia (yellow), which appears first with Structure's K=10, and later in both analyses, with K=13 and K=14. The contents of this cluster fluctuate to some extent; with Structure's K=10 and K=13 it covers Border Karelia (eastern parts), while in other cases it does not. Some clusters also only appear with one type of analysis: a cluster covering South Savo + Savonlinna transitional (light blue) appears only in Structure, with K=12 and K=14; similarly, Central Ostrobothnia (turquoise) only appears with Structure (K=13), as does Border Karelia (aniline red), with K=14. One unusual cluster, tentatively pointed out in section 4.2 and also marked with aniline red, appears with K-medoids with K values between 9 and 14. This covers Border Karelia, Ingria and a small selection across the border areas. Upon closer inspection the municipalities covered by this cluster are among those that are less extensively documented by the dialect atlas, suggesting that K-medoids is more sensitive to how well-documented a data point is than Structure.

One area that deserves further attention is the dark red cluster, covering areas in Central Finland, Kainuu and the Savonian Wedge. These appear as one large (geographically discontinuous) cluster with K=11 in K-medoids, while the Structure analyses point more towards a strong transition in this area, with either end varyingly serving as the core area; Structure's K=9 and K=13 shows Central Finland as the core, and the rest (K=11, K=12, K=14) show Kainuu as the core. With higher K values K-medoids also distinguishes the core areas as separate clusters. In these cases Central Finland is colored gray and Kainuu dark red on the maps.

Finally, we should also note that the fluctuation in the eastern area is reflected in the shape of the green cluster identified as Savo with lower K values, which reduces to either East Savo (K=9, K=11 and K=13 in Structure, K=11-14 in K-medoids) or North Karelia + North Savo (K=12 and K=14 in Structure).

(Fig. 8 here)

4.3. Stability of the results

4.3.1 Comparing different diploid runs with CLUMPP

The results in Figs 7-8 only show the Structure runs with the highest likelihoods. However, because of the stochastic nature of Structure the results can vary across runs even when the same K value and general parameters are used. In some cases we may see *label switching* – i.e. the different repetitions with the same K value identify same clusters but show them in a different order in Structure's results. The different runs may also reveal *genuine multimodality*, where independent runs of the same K value produce qualitatively different clusterings.

To overcome the problem of needing to choose a single Structure run to represent the results of a K value, the results of multiple runs can also be combined using a tool called CLUMPP. The tool takes Structure's results, aligns the populations from the analyses run with the same K value so that they match each other as well as possible (solving possible label switching problems), and produces a combined result from the membership coefficients of all the runs of the same K value. Consequently, cases where the clusters match each other well in different repetitions of K appear similar to how they are shown in the highest likelihood run, whereas areas where the inferred populations differ across repetitions become more ambiguous and show more transitions across the populations.

Differences between the CLUMPP visualizations (Fig. 9) and the highest likelihood runs (Figs 7-8) show that repetitions with same K value do not always identify the same clusters as the highest likelihood run. This shows up as more admixed populations in the CLUMPP visualizations compared to the highest likelihood runs. For instance, the Southwest dialects (light red in Fig 9) as well as the south of Häme (light purple in Fig 9) appear as more admixed areas with K=3 than they did with the highest likelihood runs (red and purple in Figs 7-8), suggesting that some of the repetitions identified different populations. Based on the pattern that the admixture shows, the conflicting populations might be somewhat closer to the results of the K-medoids analyses. In K=4 the Southwest dialects become more coherent (red in Fig 9), indicating that the independent repetitions agree on that area, whereas Häme and the lower part of the eastern dialects are more ambiguous, showing light purple and light green areas which essentially suggest that some of the repetitions with K=4 identified the Southeast dialects rather than Häme.

Some of the highest likelihood Structure populations agree well with the CLUMPP visualizations, indicating that they are quite stable, e.g. the eastern and western dialects. There are also some dialect areas that do not appear directly in the CLUMPP visualizations, although they were present in the highest likelihood runs, such as Central Ostrobothnian (turquoise) and Central Karelia (yellow). Also, the Southwest transitional area (light green) appears only weakly with K=14. This suggests that these areas are not necessarily as robustly supported by the variation data from the atlas as the other dialectal areas.

(Fig. 9 here)

4.3.2. Comparing Structure diploid, Structure haploid and K-medoids

Looking further into the stability of the dialect clusters inferred in the analyses, we compared the K-medoids results with the diploid results, as well as the haploid results (a more extensive comparison of these can be seen in the appendices). Despite differences between the analysis methods, differences in data representation, and the smaller amount of represented dialectal variation in the haploid data compared to the diploid data, the results from these three are surprisingly close to one another with lower K values (2-8). With these values all the approaches classify the dialects identically, with the exception of K=3, which is in agreement only between the two Structure runs. With higher K values the results begin to disagree somewhat more, as was already seen in section 4.2.2.

4.4. Testing for linked features

As was mentioned in section 3.3.3, Structure assumes that the data is in 'linkage equilibrium', meaning that it should cover only independent loci (or in our case uncorrelated linguistic features). Although we explored the data as a whole, we also tested for possible connectedness of the features using an approach based on counting the number of municipality pairs with identical dialect features across map pages (see section 3.3.3 for a more detailed explanation). The heat map produced from this comparison is given in Fig. 10.

(Fig. 10 here)

The histogram indicates that connectedness between the linguistic features is generally modest or low (yellow-orange). Potential problems with correlating features lie in the features located in the right side (red). As a whole the map pages do not seem to be extensively linked to one another; there are some map pages that produce somewhat higher linkage estimates than others, e.g. the map pages 70-80. However, upon closer inspection, some of these higher linkage estimates appear to be focused on maps covering smaller geographical areas. As in these cases the linkage test discards municipalities without any recorded dialect features, the estimates are based on a smaller number of municipality pairs. These cases are identifiable from the results by examining how many potential linkage (Lp) and actual linkage (La) cases were recorded when comparing the pages, and filtering out cases whose Lp or La values are below some threshold (the appendices include an example heat map filtered in this way). Notably, the linkage test in its current form seems to produce biased results on map pages which contrast a dialect variant with a very small geographical area with another variant covering the remainder of the map. Such a case can be seen for instance on page 137, with contrasts a characteristic limited to a particular area within the Eastern dialects with a characteristic covering the rest of the map; this shows up as a reddish stripe at the corresponding position on the heat map, suggesting systematic linkage of this map page to all the other pages.

5. Discussion

Our results suggest that Structure and similar population genetic clustering tools could be of value for linguists investigating intra-lingual data once it has been appropriately formatted; population genetic clustering inferred Finnish dialect areas quite sensibly. Below we sum up the results and the restrictions of the analyses, and describe how the results could serve as a basis for further analyses.

5.1 Individual dialect divisions vs. principal dialect areas of Finnish

The focus of this paper is generally less on amending the Finnish dialect division and more on exploring the suitability of new methodology for modeling dialectal diversity. Here we briefly look at the results against the principal dialect divisions outlined in the literature (2, 3 or 4 dialect areas), and also see how the division into 8 dialect areas compares with the accustomed eight dialect areas (see Itkonen 1964).

The ΔK values suggested the two-way division (K=2) as the best uppermost hierarchal division. The resulting division is a fairly accurate match with traditional Eastern and Western dialect areas. The K=2 results were generally very uniform, with the different analyses showing only minute differences along the borders. Structure's results were more descriptive, also showing the transitional areas along the border. The two-way split also remained fairly uniform with the haploid analysis and the stability of the diploid east-west division can also be seen in CLUMPP visualization. Considering the emphasis on morphophonological features in Kettunen's atlas, this was also the expected result.

Three-way divisions had a slightly higher likelihoods than the two-way division with Structure but were not supported by average silhouette values or ΔK values. Different analyses disagreed to some extent on how the data should be divided with K=3. Interestingly, the suggestions line up fairly well with divisions suggested in the literature. The K-medoids result was a fairly close match to the three-way division from Leino *et al.* (2006) and Hyvönen *et al.* (2007). Structure's highest likelihood, on the other hand, seemed to follow the three-way division originating from Lenqvist, later discussed in Paunonen (1991; 2006) and Mielikäinen (1991), with Southwest standing out as a principal dialect area. CLUMPP visualization for K=3 showed fuzzier populations especially around Häme, suggesting disagreement between the Structure runs.

The four-way division in the present analyses differed from Paunonen's (2006) four-way division, with South Ostrobothnia grouped together with Middle / North Ostrobothnia dialects and Far North, and not with South Western transitional dialects and Häme dialects as in Paunonen. Also here Structure's likelihood was slightly higher but the division was otherwise not highly supported.

One notable difference between the K=8 division when compared against the customary eight-way dialect division of Finnish is absence of the Southwest transitional dialect area (which only appeared with higher K values), and its replacement with Southeast Häme (see e.g. Rapola 1969, Wiik 2004). This division is compatible with the east-west dichotomy, with six western and two eastern clusters, emphasizing how strongly the two-way division is rooted in the data.

An interesting experiment for future studies might be to subdivide the populations produced by K=2 separately, as is done for a different type of data in Evanno *et al.* (2005). With this we could also attempt to explore the robustness of the traditional eight-way division. This would make sense since the traditional eight-way division of Finnish is subordinate to the two-way division.

5.2 Inferred dialect clusters compared with existing knowledge of Finnish dialect areas

In addition to looking at divisions with specific K values, we also visually compared our inferred dialect clusters with dialect division maps found in the literature, including Kettunen (1940a), Hakulinen (1950), Rapola (1969), Hormia (1978), Mielikäinen (1994), Savijärvi & Yli-Luukko (1994), Itkonen (1964; 1989),[6] as well as some of the dialectometrical maps from Wiik (2004). This was done by scanning the maps from the literature and scaling them so that they matched our dialect maps with clusters as well as possible. Overlapping the images of the maps provided a fairly straightforward way to compare not only our results against traditional dialect maps from the literature but also our results produced with different data sets against each other.

By comparing the highest likelihood visualizations of the Structure analyses and the visualizations of the K-medoids analyses against each another, and also inspecting the runs with different K values against each other, we could group the clusters into 38 distinct dialect areas with specific borders. 20 of these dialect areas matched closely with the dialect divisions shown in the literature. Structure's highest likelihood analyses across different K values covered 18 of these attested dialect areas, and K-medoids covered 17.

Some general trends were also apparent from the visual comparisons. For instance, the western dialect areas were notably much more stable and coherent across the analyses while the eastern dialects fluctuated more. This could reflect the differences in the histories of western and eastern Finns: in the east the gradual expansion to north and their slash and burn agriculture made them more mobile than the people in the west, who had more stable settlements and land ownership (Virrankoski 2012). The ambiguity in the east could also reflect the relatively young nature of the dialects in this area. For instance, the Savo dialects emerged according to Wiik (2004) only around 1000 years ago and their gradual expansion and mixing with other dialects was still going on around 300 years ago. In any case, an in-depth look into the dialect transitions could be of interest in the future.

5.3. Using population genetic clustering as a basis for further study

The correspondence between the results of the different methods along with good agreement against traditional dialect areas suggest that dialect data can be examined successfully with population genetic tools. Proper verification of the approach creates a solid basis for future applications of microevolutionary methodology, which provide a huge potential for shedding light on linguistic phenomena. However, the present study has only scratched the surface of applying population genetic tools for exploring dialect material. Population genetics provides a framework and the tools to examine language-internal variation, such as how the linguistic variation is (spatially) organized or why the dialects emerged and how they are maintained. Indeed, Wieling & Nerbonne (2015) call

[6] Most of these maps can be found in Wiik (2004).

for studies focusing on resolving factors underlying linguistic variation. We hope that the approach presented in this paper contributes to achieve this goal.

The focus of this paper is on thoroughly exploring the cornerstone of many population genetic analyses – clustering the data as populations – but besides of this we shortly present some examples of the possibilities the model-based clustering methods provide for further studies. In the methods section we have already shown that the results themselves can be visualized in a variety of ways according to the needs of the study. Here, we give two examples of how the membership coefficients can be used for calculating new measurements.

The membership coefficients include information that reflects population admixture, and these values can be used to quantify how diverse the language of a given municipality is. This diversity may be calculated for example with Shannon-Wiener diversity index (H) (also called e.g. as Shannon's entropy) (Legendre & Legendre 2012) which is one of the diversity indices commonly used in ecology to measure the diversity of ecological communities. In general, it uses proportions of characters of interest to calculate the diversity, for which we can use the inferred IC values. It is calculated by multiplying each IC value of a given municipality with its logarithm, and by taking the negative sum of these, i.e.

$$H \quad = -\sum_{i=1}^{q} p_i \log(p_i)$$

In the case of languages the index is low when the amount of linguistic diversity is low, that is, when traits specific to one dialect only is dominant in a certain municipality (Figure 11). In contrast, the index is high when the municipality harbors traits typical to multiple dialects and the dialects are present in equal frequencies. Diversity values could be e.g. further compared to other spatial attributes to understand why the high linguistic diversity is located in the given areas.

(Fig. 11 here)

The similarity of the inferred populations in relation to one another is another thing that we cannot directly see from Structure's membership coefficients; Structure essentially produces populations with which it can describe the entirety of the data, but we cannot directly see how (linguistically or genetically) similar or different these populations actually are. Quantifying the linguistic differences between the inferred dialect populations could be a topic of interest for example to study drivers of dialectal divergence (which is the focus in our forthcoming paper, Honkola *et al.* ms). A population genetic metric that allows us to shed light on this is F_{ST}, which estimates the amount of genetic differentiation between populations. A related metric Φ_{ST}, has been used outside of biology for studying differences between folktale types (Ross *et al.* 2013).

In principle F_{ST} measures how much reduction there has been in heterozygosity (i.e. changes in the allele frequencies) due to subpopulation divergence (Hamilton 2009). Thus, it compares the total expected heterozygosity of all the populations (H_T) with the averaged expected heterozygosity of the studied subpopulations (H_S), and is calculated with the following formula:

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

If the expected averaged heterozygosities of the subpopulations equal to the total expected heterozygosity of all the populations, ($H_T = H_S$), the allele frequencies in the different subpopulations do not differ from each other and suggests that there is no population structure. However, if these values differ, it suggests that the population has an inferrable substructuring.[7] Further, large F_{ST} values indicate large differences in the allele frequencies of the populations and thus greater differentiation, while small values reflect more similar allele frequencies and consequently smaller differences between populations.

The expected heterozygosities needed for the F_{ST} calculation are calculated from the observed allele frequencies in a population. Therefore it is possible to calculate the expected 'linguistic heterozygosities' from the observed frequencies of different linguistic variants. For the example analysis shown here, we isolated the "core areas" from the results of a K=14 analysis, i.e. municipalities with IC values > 0.75 (Fig. 12). We then calculated the linguistic differences between the core areas with F_{ST}. The F_{ST} values shown here were calculated using GenAlEx (Peakall & Smouse 2006, 2012), but there are many other tools available.

The F_{ST} values varied between 0.81 and 0.42 (Table 1), reflecting stronger differences than one tends to find with comparable biological data. It is likely that the nature of the dialect data affects this; unlike genetic data, which covers a systematic sample collected without an intention of maximizing population-wise differences, the dialect atlas covers features that serve to highlight the contrasting characteristics of Finnish dialects as well as possible. From a linguistic perspective our F_{ST} values are generally distributed as one would expect them to; for instance, the six highest pairwise F_{ST} values are all according to the east-west dichotomy. The six lowest F_{ST} values indicate similarity between 1) Savo, Karelia and Southeast dialects, 2) the Häme dialects and the Southwestern transitional dialects, and 3) Kainuu dialects and Middle / North Ostrobothnia,[8] all of which are plausible transitional areas (cf. e.g. Wiik 2004). We have used these values elsewhere and compared the dialectal differences to cultural, environmental and administrative differences of the same areas (Honkola *et al.* ms.).

(Fig. 12 here)
(Table 1 here)

[7] Inferring population structure from the differences of heterozygosities is based on the idea that if there is subpopulation structure the subpopulations differ in their allele frequencies and their averaged heterozygosity cannot be as high as in the total population.

[8] The low F_{ST} values between Kainuu and Central/North Ostrobothnia reflect what is shown on map 14 in Wiik (2004) – that is, the border between the eastern and the western dialect areas is the least steep around this area. Notably, Hyvönen *et al.* (2007) also produced a combined cluster of Kainuu and Central/North Ostrobothnia using lexical data, further highlighting the fuzziness of the east-west border in this area.

6. Conclusion

In this paper we endeavored to take Finnish dialect studies to a less explored methodological direction, examining the dialect atlas of Finnish with population-genetic and distance-based clustering. We have had the fortunate situation of being able to take advantage of the vast knowledge that exists on Finnish dialects, which has also allowed us to focus more on methodological matters. We did not dig extremely deep into the intricacies of the Finnish dialect division, and instead focused more on exploring new approaches for analyzing the Dialect Atlas of Finnish and discussing analogies between within language and within-species variation, which in many ways is at the core of this approach.

The results suggest that population genetic clustering performs reasonably well with dialect data. In general the clusterings did not significantly clash with existing dialect research, and although biological allele and dialect datasets have notable differences, population genetic clustering was able to capture dialect variation quite well. The different analyses produced fairly consistent results especially with lower K values. Although the traditional K-medoids clustering was also quite efficient in inferring dialect clusters, a clear advantage for Structure were there resulting membership coefficient values, which allow for detailed visualization (e.g. soft clusters) and further research and allowing to take into account the missing data in the analyses. This also allows one to explore the proportion of admixture as meticulously as one desires.

The expectations built into biological analysis tools and their potential effects on the results are an important matter to consider when dealing with non-biological data. Firstly, Structure's algorithm models populations as sets of allele frequencies that are compared to the allele frequencies of the "model population", with allele frequencies in Hardy-Weinberg equilibrium (HWE). This does not mean that the input data needs to be in HWE. Therefore, HWE does not truly limit what we can study, but can limit how we can interpret the results. If the object of our study is not in HWE, its allele frequencies are undergoing a change although the analysis assumes them to remain the same. In this case we run into problems if we assume that the populations that Structure infers are accurate representations of populations much more ancient than the data we have analyzed, especially if the inferred population is small. Secondly, for all statistical analyses the variables should be independent from each other; in the case of Structure it assumes that each of the loci (dialectal feature) should be independent from the other loci. We could not find a test for feature-wise linkage for the kind of linguistic data that we have that could be readily adopted; this prompted us to devise a simple preliminary test for this purpose, which did not point to significant linkage in the dialect atlas. However, as the method we used is a preliminary metric that has not been extensively tested, the matter of linkage should be given further attention in future studies.

Computational approaches in historical linguistics, such as phylogenetics, have initiated a field in linguistics essentially analogous to the study of macroevolution of biological species. The population genetic framework, which operates on a microevolutionary level, could be used to study variation within a language. Here, we have adopted this approach for dialect study and present some possible applications, in the hope that this approach could open new doors for studying linguistic variation in the future.

Bibliography

Agricola, Michael. 1548. *Se Wsi Testamenti.* Stockholm: Amund Lauritzon. Facsimile edition 1987. Helsinki: WSOY.

Beaumont, Mark A. & Bruce Rannala. 2004. "The Bayesian revolution in genetics". *Nature Reviews Genetics* 5: 251–261.

Bowern, Claire. 2012. "The riddle of Tasmanian languages." *Proc. R. Soc. B*: 2012 -; DOI: 10.1098/rspb.2012.1842. Published 26 September 2012.

Campbell, Neil A., Jane B. Reece, Lisa A. Urry, Michael L. Cain, Steven A. Wasserman, Peter V. Minorsky & Robert B. Jackson. 2008. *Biology.* 8th ed. Boston: Pearson Benjamin Cummings.

Chambers, J. K. & Peter Trudgill. 1998. *Dialectology*. 2nd ed. Cambridge: Cambridge University Press.

Chen, Chibiao, Eric Durand, Florence Forbes and Olivier François. 2007. "Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study". *Molecular Ecology Notes* 7: 747-756.

Corander, Jukka, Partik Waldmann and Mikko J. Sillanpää. 2003. "Bayesian Analysis of Genetic Differentiation Between Populations". *Genetics* 163, 367-374.

Croft, William. 2000. *Explaining language change: An evolutionary approach.* Essex: Pearson Education Limited.

Earl, Dent A. & vonHoldt, Bridgett M. 2012. "STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method." *Conservation Genetics Resources* 4(2): 359-361.

Dunn, Michael, Stephen C. Levinson, Eva Lindström, Ger Reesink & Angela Terrill. 2008. "Structural phylogeny in historical linguistics: Methodological explorations applied in Island Melanesia." *Language* 84(4): 710-759.

Embleton, Sheila & Eric S. Wheeler. 1997. "Finnish dialect atlas for quantitative studies". *Journal of Quantitative Linguistics* 4:1-3, 99-102.

Embleton, Sheila & Eric S. Wheeler. 2000. "Computerized dialect atlas of Finnish: Dealing with ambiguity". *Journal of Quantitative Linguistics* 7:3, 227-231.

Evanno, Guillaume, Sebastien Regnaut & Jérôme Goudet. 2005. "Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study." *Molecular Ecology* 14: 2611-2620.

Francis, Roy M. 2014*. pophelper: An R package for analysis of STRUCTURE and TESS files.* R package version 1.0.0.

Gasmi, Laila, Helene Boulain, Jeremy Gauthier, Aurelie Hua-Van, Karine Musset, Agata K. Jakubowska, Jean-Marc Aury, Anne-Nathalie Volkoff, Elisabeth Huguet, Salvador Herrero & Jean-Michel Drezen. 2015. "Recurrent Domestication by Lepidoptera of Genes from Their Parasites Mediated by Bracoviruses." *PLOS Genetics* (DOI:10.1371/journal.pgen.1005470)

Geographical Society of Finland. 1928. *Atlas of Finland 1925.* Helsinki: Otava.

Gray, Russell D. & Quentin D. Atkinson. 2003. "Language-tree divergence times support the Anatolian theory of Indo-European origin". *Nature* 426: 435–439.

Hakulinen, Lauri. 1950. "Kansankielen sanakirjan koeartikkeleja." *Virittäjä* 54: 425-444.

Hamilton, Matthew. 2009. *Population Genetics*. Oxford: Wiley-Blackwell.

Harju, Erkki-Sakari (ed.) 2009. *Suomen kartta 1920*. Facsimile edition. Helsinki: Karttakeskus.

Heeringa, Wilbert. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance.* Groningen dissertations in Linguistics 46. Groningen: University of Groningen.

Honkola, Terhi, Outi Vesakoski, Kalle Korhonen, Jyri Lehtinen, Kaj Syrjänen & Niklas Wahlberg. 2013. "Cultural and climatic changes shape the evolutionary history of the Uralic languages". *Journal of Evolutionary Biology* 26: 1244–1253.

Honkola, Terhi, Kalle Ruokolainen, Kaj Syrjänen, Antti Leino, Ilpo Tammi, Niklas Wahlberg & Outi Vesakoski. (ms). "Evolution within a language: environmental differences contribute to dialect diversification."

Hormia, Osmo. 1978. *Finska dialekter: en översikt*. Lund: Liberläromedel.

Hovdhaugen, Even, Fred Karlsson, Caron Henriksen & Bengt Sigurd. 2000. *The History of Linguistics in the Nordic Countries.* Helsinki: Societas Scientiarum Fennica.

Hurtta, Heikki. 1999. "Variaationtutkimuksen myytit ja stereotypiat". In: *Kirjoituksia Sosiolingvistiikasta* ed. by Urho Määttä, Pekka Pälli & Matti K. Suojanen. Tampere: University of Tampere.

Hyvönen, Saara, Antti Leino & Marko Salmenkivi. 2007. "Multivariate Analysis of Finnish Dialect Data – An Overview of Lexical Variation". *Literary and Linguistic Computing* 22: 3, 271–290.

Itkonen, Terho. 1964. *Proto-Finnic Final Consonants. Their History in the Finnic Languages with Particular Reference to the Finnish Dialects. I: 1 Introduction. The history of k in Finnish.* Helsinki: Suomalaisen Kirjallisuuden Kirjapaino.

Itkonen, Terho. 1989. *Nurmijärven murrekirja*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Jakobsson, Mattias & Noah A. Rosenberg. 2007. "CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure." *Bioinformatics* 23:14, 1801-1806.

Kaufman, Leonard and Peter Rousseeuw. 1987. "Clustering by means of Medoids." In: *Statistical Data Analysis Based on the L1-Norm and Related Methods* ed. by Yadolah Dodge, pp. 405-416. Amsterdam: North-Holland.

Kettunen, Lauri. 1930. *Suomen Murteet II. Murrealueet*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Kettunen, Lauri. 1940a. *Suomen Murteet III A. Murrekartasto.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Kettunen, Lauri. 1940b. *Suomen Murteet III B. Selityksiä Murrekartastoon.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Kettunen, Lauri. 1960. *Matkapakinoita ja muita muistelmia: 1925-1960.* Helsinki: Kauppakirjapaino Oy.

Kriegel, Hans-Peter, Peer Kröger & Arthur Zimek. 2009. "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering". *ACM Transactions on Knowledge Discovery from Data* 3:1, article 1.

Laurosela, Jussi. 1922. *Foneettinen tutkimus Etelä-Pohjanmaan murteesta*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Lee, Sean & Toshikazu Hasegawa. 2011. "Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages". *Proceedings of the Royal Society B*. 278: 3662–3669.

Legendre, Pierre & Louis Legendre. 2012. *Numerical Ecology*. *Third English Edition.* Amsterdam, Oxford: Elsevier.

Lehtinen, Jyri, Terhi Honkola, Kalle Korhonen, Kaj Syrjänen, Niklas Wahlberg & Outi Vesakoski. 2014. “Behind Family Trees: Secondary Connections in Uralic Language Networks”. *Language Dynamics and Change* 4: 189-221.

Leino, Antti & Saara Hyvönen. 2008. "Comparison of Component Models in Analysing the Distribution of Dialectal Features". *International Journal of Humanities and Arts Computing* 2, 173–187.

Leino, Antti, Saara Hyvönen & Marko Salmenkivi. 2006. ”Mitä murteita suomessa onkaan? Murresanaston levikin kvantitatiivista analyysiä”. *Virittäjä* 110, 26–45.

Leskinen, Heikki. 1992. *Karjalan kielikartasto 1. Idän ja lännen sanastoeroja.* Jyväskylä: Jyväskylän yliopiston suomen kielen ja viestinnän julkaisuja 21.

Levinson, Stephen C. & Russell D. Gray. 2012. "Tools from evolutionary biology shed new light on the diversification of languages." *Trends in Cognitive Sciences* 16(3), 167-173.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. 2014. *cluster: Cluster Analysis Basics and Extensions.* R package version 1.15.3.

Mielikäinen, Aila. 1991. *Murteiden murros. Levikkikarttoja nykypuhekielen piirteistä.* Jyväskylä: Jyväskylän yliopisto.

Mielikäinen, Aila. 1994. *Etelä-Savon murteiden äännehistoria II. Vokaalit.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Nerbonne, John & William Kretzschmar. 2003. ”Introducing Computational Techniques in Dialectometry”. *Computers and Humanities* 37: 245-255.

Pagel, Mark. 2009. ”Human language as a culturally transmitted replicator.” *Nature Reviews Genetics* 10: 405-415.

Palander, Marjatta. 1999. ”Mitä dialektometria on?” *Virittäjä* 103, 259-265.

Palander, Marjatta, Lisa-Lena Opas-Hänninen & Fiona Tweedie. 2003. ”Neighbours or Enemies? Competing Variants Causing Differences in Transitional Dialects”. *Computers and the Humanities* 37:3, 359-372.

Paunonen, Heikki. 2006. ”Lounaismurteiden asema suomen murteiden ryhmityksessä”. In: Nordlund, Taru, Tiina-Onikki Rantajääskö & Toni Suutari (eds.) *Kohtauspaikkana kieli – näkökulmia persoonaan, muutoksiin ja valintoihin*, pp. 249–268. Helsinki: Suomalaisen Kirjallisuuden Seura.

Paunonen, Heikki. 1991. ”Till en ny indelning av de finska dialekterna”. *Fenno-Ugrica Suecana* 10, 75-79.

Peakall, Rod & Peter E. Smouse. 2012. “GenAlEx 6.5: genetic analysis in Excel. Population genetic software for teaching and research-an update.” *Bioinformatics* 28, 2537-2539.

Peakall, Rod & Peter E. Smouse. 2006. “GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes*. 6, 288-295.

Pritchard, Jonathan K., Matthew Stephens & Peter Donnelly. 2000. ”Inference of Population Structure Using Multilocus Genotype Data.” *Genetics* 155: 945-959.

Pritchard, Jonathan K., Xiaoquan Wen & Daniel Falush. 2010. *Documentation for structure software: version 2.3*.

R Core Team. 2014. *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. URL http://www.R-project.org/.

Rapola, Matti. 1969. *Johdatus suomen murteisiin.* 3rd ed. Helsinki: Suomalaisen Kirjallisuuden Seura.

Reesink, Ger, Ruth Singer & Michael Dunn. 2009. “Explaining the Linguistic Diversity of Sahul Using Population Models.” PLoS Biology 7:11.

Ross, Robert M., Simon J. Greenhill & Quentin D. Atkinson. 2013. “Population structure and cultural geography of a folktale in Europe.” *Proceedings of the Royal Society B*. 280: 20123065.

Rousseeuw, Peter J. 1986. ”Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.” *Journal of Computational and Applied Mathematics* 20: 53-65.

Sarmela, Matti (ed.) 1994. *Atlas of Finnish ethnic culture 2. Folklore.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Savijärvi, Ilkka & Eeva Yli-Luukko. 1994. *Jämsän äijän murrekirja.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Syrjänen, Kaj, Terhi Honkola, Kalle Korhonen, Jyri Lehtinen, Outi Vesakoski & Niklas Wahlberg. 2013. “Shedding more light on language classification using basic vocabularies and phylogenetic methods. A case study of Uralic.” *Diachronica* 30:3: 323-352.

Tuomi, Tuomo (ed.). 1989. *Suomen murteiden sanakirja. Johdanto.* Helsinki: Kotimaisten kielten tutkimuskeskus.

Vhaël, Bartholdus G. 1733. *Grammatica fennica.* Åbo: Johan Kiämpe.

Virrankoski, Pentti. 2012. *Suomen historia: Maa ja kansa kautta aikojen*. Hämeenlinna: Suomalaisen Kirjallisuuden Seura.

Vuorela, Toivo (ed.). 1976. *Atlas of Finnish Folk Culture 1. Material culture.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Warelius, Anders. 1848. ”Bidrag till Finlands kännedom i ethnographiskt hänseende”. *Suomi* 7, 47-130.

Warnes, Gregory R., Ben Bolker, Lodewijk Bonebakker, Robert Gentleman, Wolfgang Huber, Andy Liaw, Thomas Lumley, Martin Maechler, Arni Magnusson, Steffen Moeller, Marc Schwartz and Bill Venables. 2014. *gplots: Various R programming tools for plotting data.* R package version 2.15.0. http://CRAN.R-project.org/package=gplots

Wieling, Martijn & John Nerbonne. 2015. “Advances in Dialectometry.” *Annual Review in Linguistics* 1:1. 243-264.

Wiik, Kalevi. 2004. *Suomen murteet. Kvantitatiivinen tutkimus.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Figure 1
The ‘gold standard’ of Finnish dialect divisions, suggested by Terho Itkonen (Itkonen 1964). The main areas are: Southwest (1a-b), Southwest transitional (2a-e), Häme (3a-f), South Ostrobothnia (4), Middle / North Ostrobothnia (5a-b), Far North (6a-e), Savo (7a-h), and Southeast (8a-c). The primary division of these dialects is between western dialects (1-6) and eastern dialects (7-8).

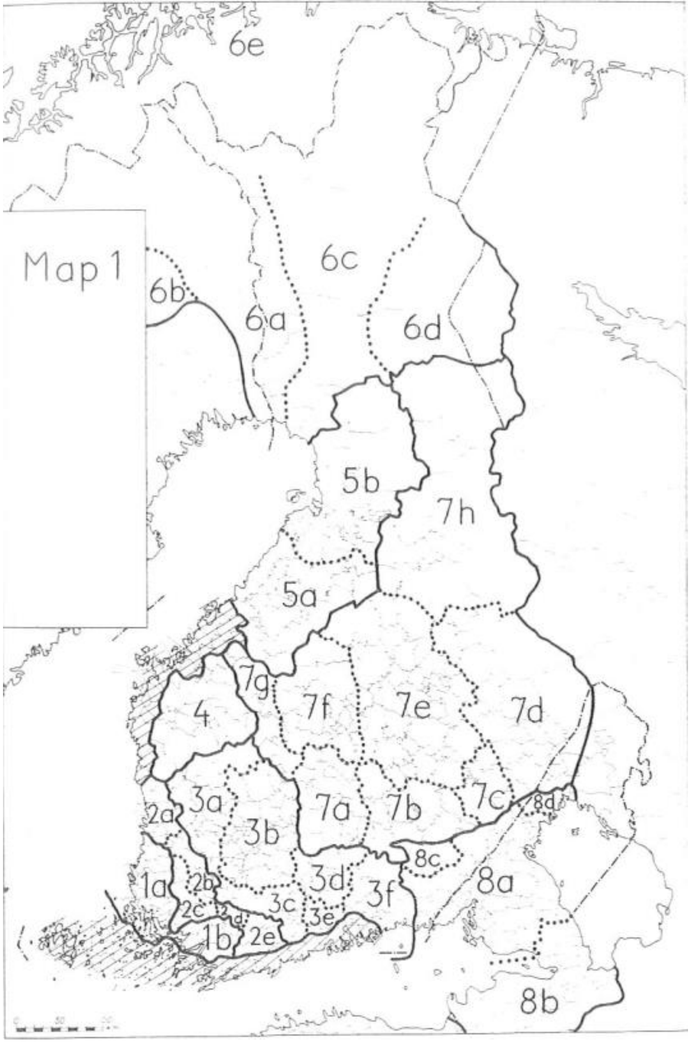

Figure 2
An example page from the Dialect Atlas of Finnish (Kettunen 1940a). The legend in the upper right lists the variants of the dialect feature that the map covers. The depicted page 8 documents morphophonological variation within the word *metsä* ('forest').

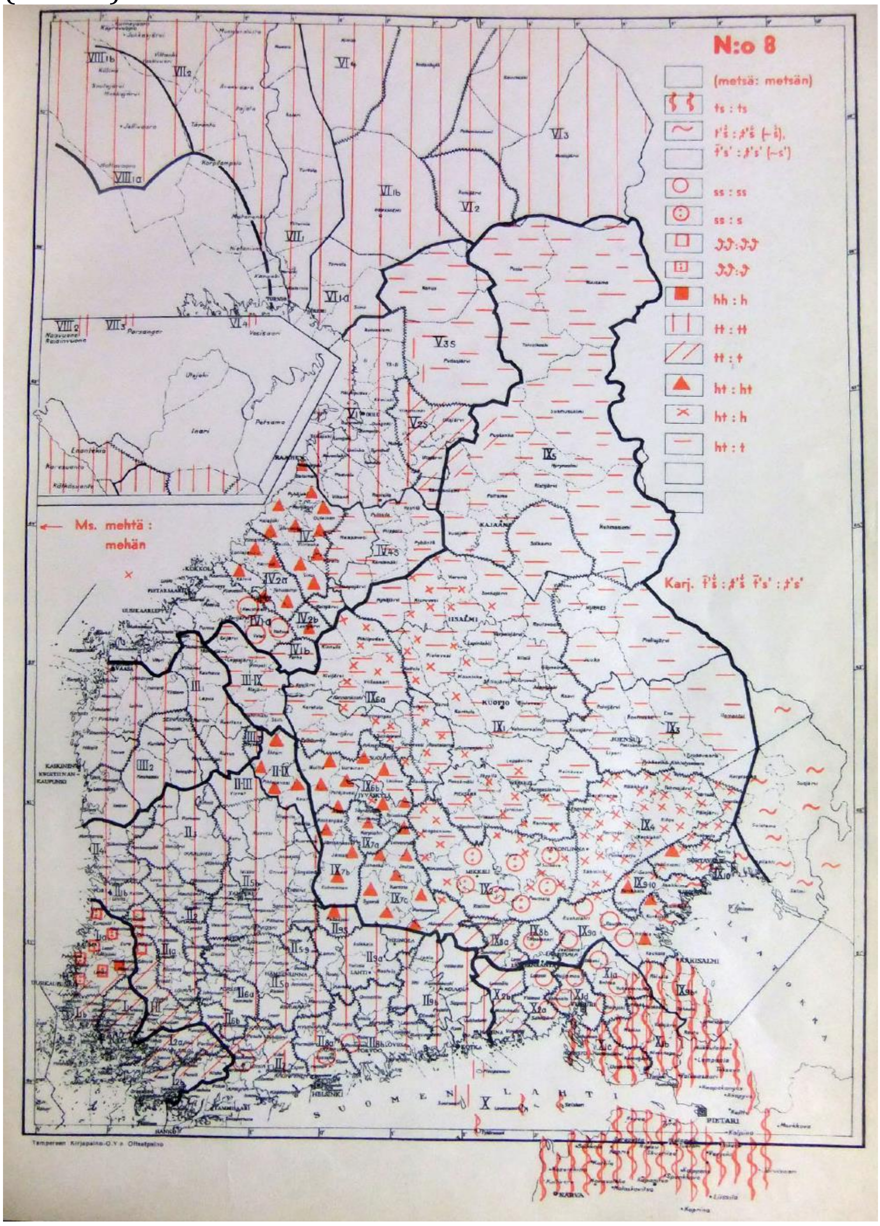

Figure 3

Two visualization styles for a division of Finnish dialects into 3 populations using Structure. Municipalities marked in white have not been studied. a) Traditional Structure barplot output. Each vertical line (perpendicular to the x-axis) represents one of the studied 525 municipalities and the color represents the dialect admixture proportions within that municipality (the frequencies of the three clusters). b) Frequency data plotted on a map, with frequencies of each inferred cluster (IC) divided to two classes: More saturated colors represent the core areas of the dialects, where the IC value is high (0.75-1); less saturated colors shows the transitional areas, with IC values between 0.5 – 0.75. c) Like b but with five frequency classes, showing the dialect transitions more accurately.

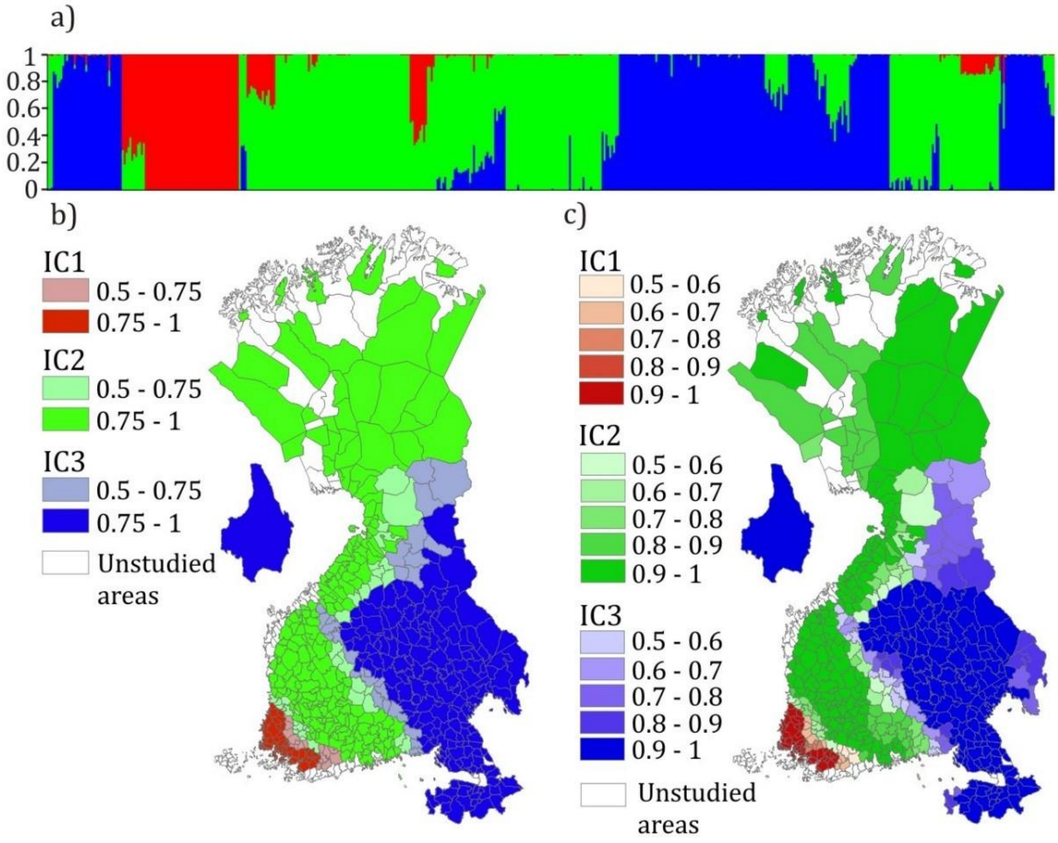

Figure 4

A close-up of South Ostrobothnia and the surrounding areas with K=8, using three visualizations: a) dialects represented with two frequency classes. Municipalities in white along the coast represent areas without data; between dialects, they represent strong admixture – i.e. all IC values below 0.5. b) The same result shown as frequency bars, revealing the dialect admixture better. c) A small part of the map with percentages also shown.

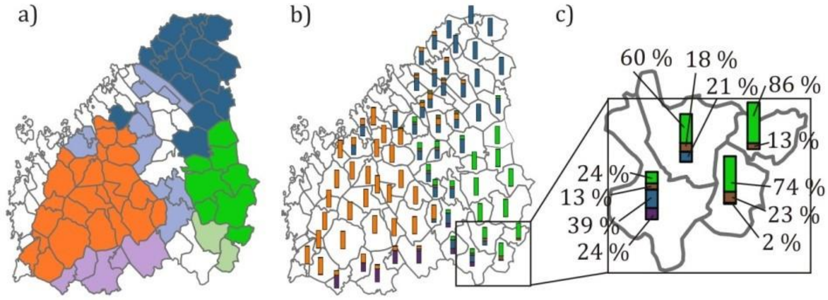

Figure 5

a) Estimated mean log likelihood of the data of K=1-20 (outliers excluded) b) *ΔK* of K=2-19 and diploid data.

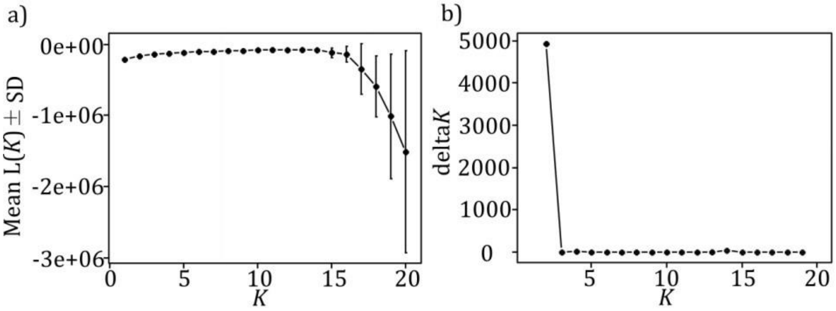

Figure 6
Average silhouette widths with *K*=2-20.

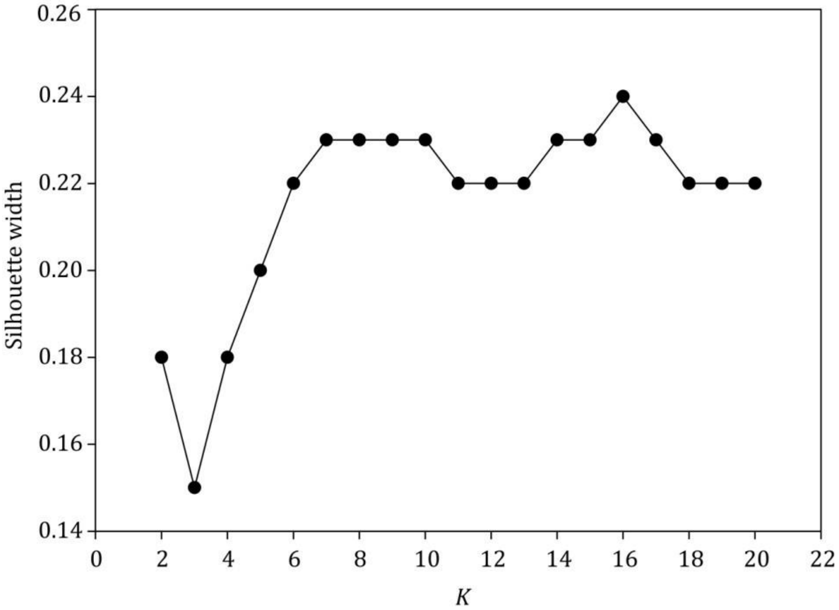

Figure 7

Dialect divisions K=2-8, with Structure diploid on the top row and K-medoids results on the bottom row. Structure diploid results use two shades of color to differentiate core areas (more saturated colors, IC values 0.75-1) from transitional areas (less saturated colors, IC values 0.5-0.75). White municipalities in the peripheral areas are undocumented whereas white municipalities in central areas indicate strong admixture (IC values under 0.50). The area shown separate from the rest of the map indicates Värmland in Sweden where people from eastern Finland migrated in the 16th century. The colors in K=8 correspond with the following dialects: red = Southwest; purple = West Häme; brown = Southeast Häme + Päijät-Häme; orange = South Ostrobothnia; blue = Middle / North Ostrobothnia + North Kainuu + Kemijoki; olive green = Far North; green = Savo; gray = Southeast. A more detailed explanation of the areas is given in section 4.2.1.

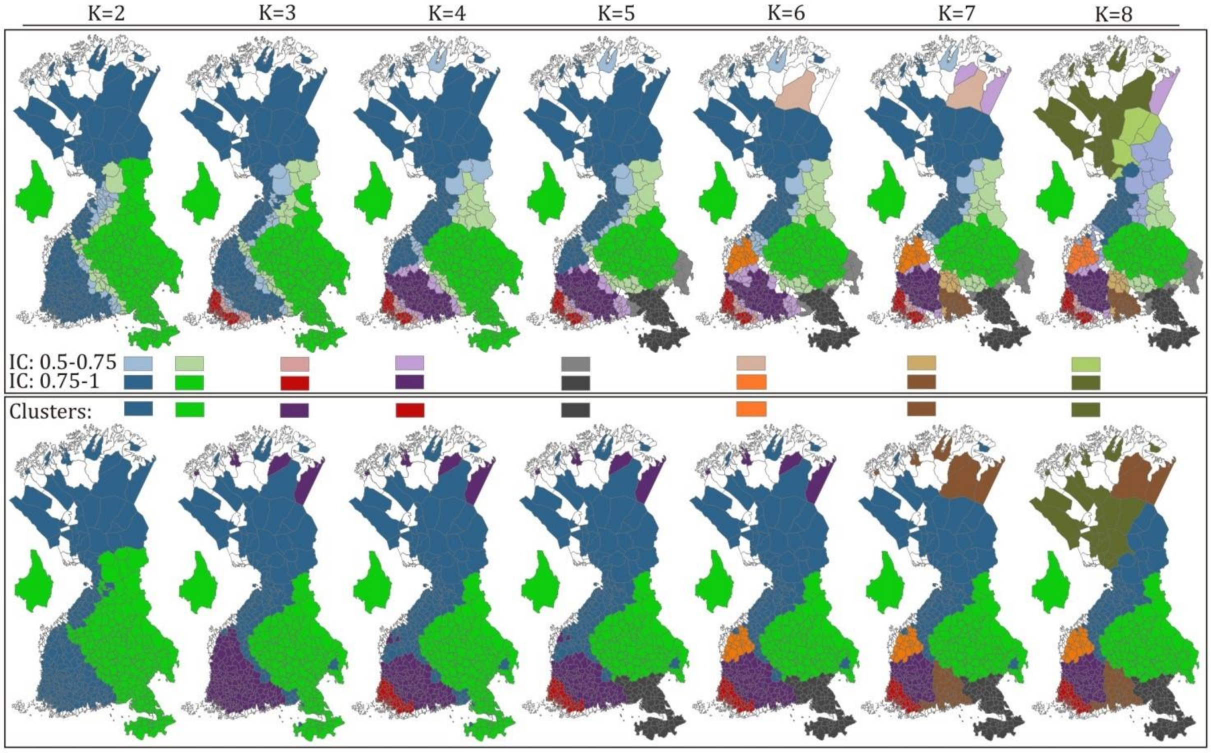

Figure 8

Dialect divisions K=9-14, with Structure diploid results presented in the top row and K-medoids results in the bottom row. The areas with the same color do not necessarily represent identical dialect areas across the maps. Other details are discussed in Figure 7. A more detailed explanation of the areas is given in section 4.2.2.

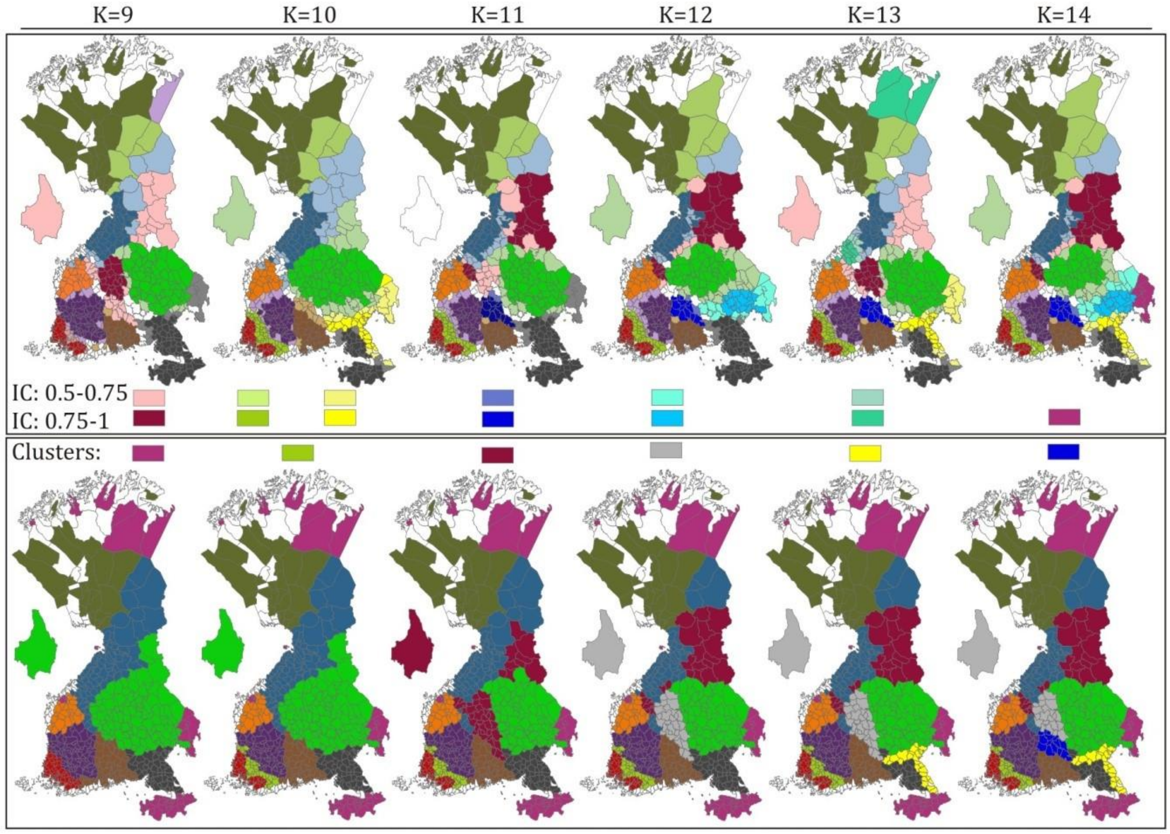

Figure 9

Dialect divisions K=2-14 visualized with CLUMPP after excluding outliers. Color pairs for municipalities below the maps are in the order of appearance to assist to observe the appearing clusters and their frequency.

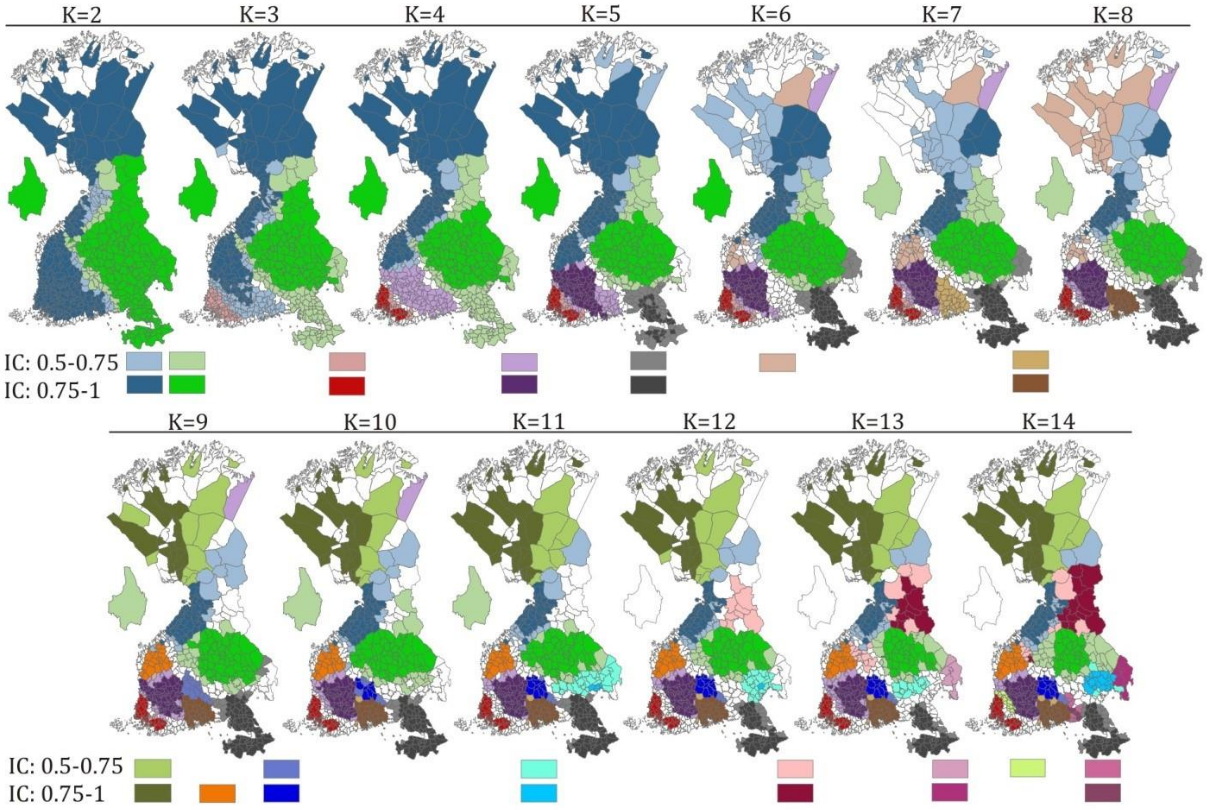

Figure 10

Heat map and histogram for the municipality pair comparisons for each map sheet. The data points along the horizontal and vertical axes correspond to the map pages of the atlas. The color scale represents the level of linkage, with red (1.0) representing a high linkage percentage, and yellow a low linkage percentage (0.0).

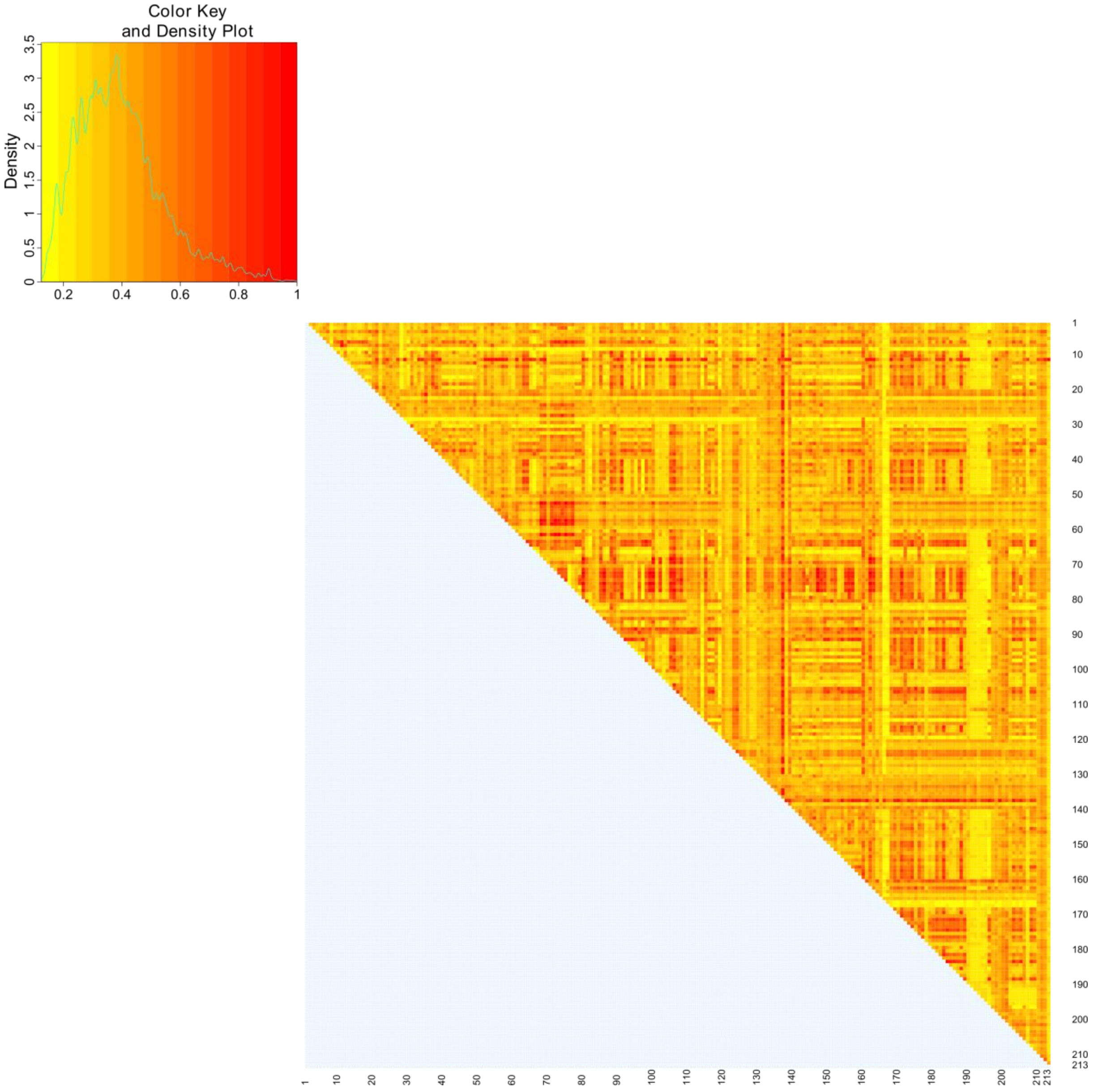

Figure 11

Shannon-Wiener indices (SWI) calculated for each municipality after dividing the data to seven populations. SWI are divided into ten equal-sized classes ranging from smallest SWI indicating lowest amount of linguistic diversity (municipalities colored with white) to the class of largest SWI indicating the largest amount of linguistic diversity (municipalities colored with black).

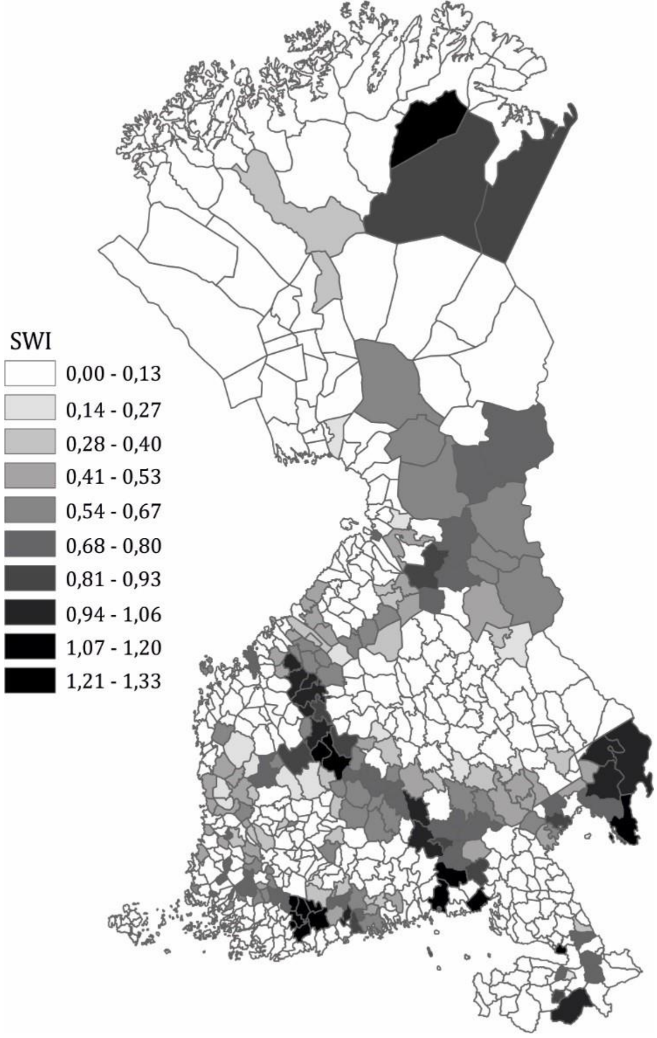

Figure 12
Core areas identified from a K=14 Structure run using an IC value threshold of 0.75.

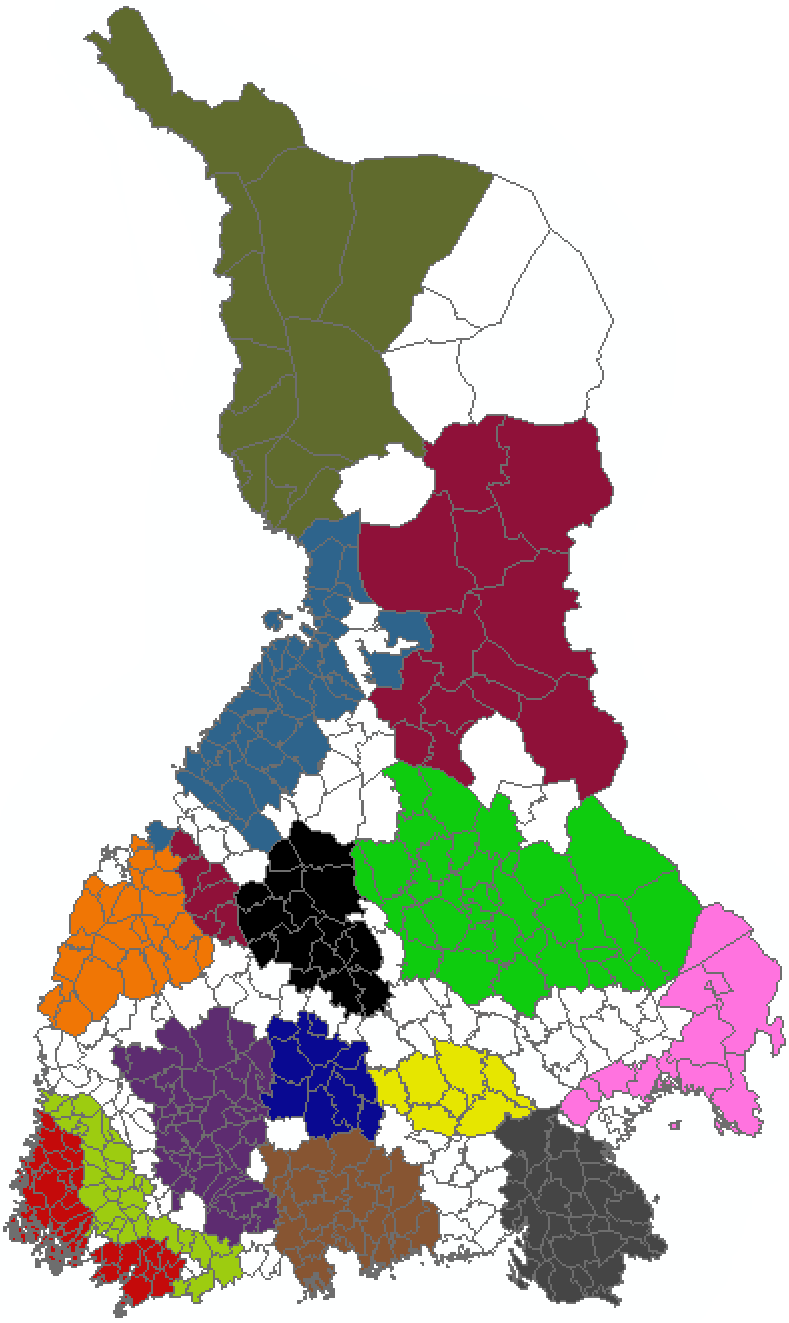

Table 1

Pairwise F_{ST} values indicating linguistic differences of the populations presented in Fig. 12. The color codes in Fig. 12 match the ones in Table 1.

	🟥	🟫	🟪	🟩	🟪	🟥	🟦	🟩	🟧	🟨	🟫	⬛	🟦	⬛
🟥	0													
🟫	0.70	0												
🟪	0.71	0.57	0											
🟩	0.50	0.55	0.62	0										
🟪	0.70	0.66	0.70	0.44	0									
🟥	0.76	0.62	0.53	0.65	0.70	0								
🟦	0.74	0.65	0.53	0.62	0.65	0.59	0							
🟩	0.81	0.76	0.55	0.75	0.78	0.58	0.69	0						
🟧	0.74	0.68	0.69	0.59	0.64	0.72	0.74	0.80	0					
🟨	0.78	0.72	0.44	0.69	0.77	0.65	0.60	0.61	0.80	0				
🟫	0.64	0.55	0.56	0.46	0.46	0.62	0.50	0.71	0.62	0.63	0			
⬛	0.81	0.78	0.60	0.73	0.78	0.57	0.66	0.54	0.81	0.67	0.69	0		
🟦	0.71	0.48	0.60	0.58	0.63	0.49	0.61	0.70	0.62	0.68	0.59	0.67	0	
⬛	0.74	0.65	0.42	0.66	0.73	0.61	0.62	0.65	0.74	0.52	0.59	0.67	0.65	0

Appendices

	Structure (diploid)	Structure (haploid)	K-medoids
K=2	Eastern	Eastern	Eastern
	Western	Western	Western
K=3	Eastern	Eastern	Eastern
	Western w/o Southwest	Western w/o Southwest	Middle / North Ostrobothnia + Far North
	Southwest	Southwest	Southwest + Häme + South Ostrobothnia
K=4	Eastern	Eastern	Eastern
	Southwest	Southwest	Southwest
	Häme	Häme	Häme
	Ostrobothnia + Far North	Ostrobothnia + Far North	Ostrobothnia + Far North
K=5	Southwest	Southwest	Southwest
	Häme	Häme	Häme
	Ostrobothnia + Far North	Ostrobothnia + Far North	Ostrobothnia + Far North
	Savo	Savo	Savo
	Southeast	Southeast	Southeast
K=6	Southwest	Southwest	Southwest
	Middle / North Ostrobothnia + Far North	Middle / North Ostrobothnia + Far North	Middle / North Ostrobothnia + Far North
	Häme	Häme	Häme
	Savo	Savo	Savo
	Southeast	Southeast	Southeast
	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
K=7	Southwest	Southwest	Southwest
	Middle / North Ostrobothnia + Far North	Middle / North Ostrobothnia + Far North	Middle / North Ostrobothnia + Far North
	Savo	Savo	Savo
	Southeast	Southeast	Southeast
	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	West Häme	West Häme	West Häme
	Southeast Häme + Päijät-Häme	Southeast Häme + Päijät-Häme	Southeast Häme + Päijät-Häme
K=8	Southwest	Southwest	Southwest
	Savo	Savo	Savo
	Southeast	Southeast	Southeast
	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	West Häme	West Häme	West Häme
	Southeast Häme + Päijät-Häme	Southeast Häme + Päijät-Häme	Southeast Häme + Päijät-Häme
	Far North	Far North	Far North
	Middle / North Ostrobothnia + North Kainuu + Kemijoki	Middle / North Ostrobothnia + North Kainuu + Kemijoki	Middle / North Ostrobothnia + North Kainuu + Kemijoki
K=9	Southwest	Southwest	Southwest
	East Savo	Savo	Savo
	Southeast	Southeast	North + Border Karelia + Ingria + Coastal
	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	West Häme	West Häme	West Häme
	Southeast Häme	Southeast Häme + Päijät-Häme	Southeast Häme
	Far North	Far North	Far North

	Structure (diploid)	Structure (haploid)	K-medoids
	Middle / North Ostrobothnia + North Kainuu + Kemijoki	Middle / North Ostrobothnia + North Kainuu + Kemijoki	Middle / North Ostrobothnia + North Kainuu + Kemijoki
	Central Finland	Southwest transitional	Southeast Proper + Savitaipale / Lemi
K=10	Savo	Savo	Savo
	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	West Häme	West Häme	West Häme
	Southeast Häme + Päijät-Häme	Southeast Häme	Southeast Häme
	Far North	Far North	Far North
	Middle / North Ostrobothnia + North Kainuu + Kemijoki	Middle / North Ostrobothnia + North Kainuu + Kemijoki	Middle / North Ostrobothnia + North Kainuu + Kemijoki
	Southwest	Southwest	Southwest
	Southwest transitional	Southwest transitional	Southwest transitional
	South Karelia	Southeast	Southeast Proper + Savitaipale / Lemi
	Central Karelia	Päijät-Häme	North + Border Karelia + Ingria + Coastal
K=11	Southeast	Southeast	Southeast Proper + Savitaipale / Lemi
	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	West Häme	West Häme	West Häme
	Southeast Häme	Southeast Häme	Southeast Häme + Päijät-Häme
	Far North	Far North	Far North
	Middle / North Ostrobothnia	Middle / North Ostrobothnia	Middle / North Ostrobothnia + North Kainuu + Kemijoki
	Southwest	Southwest	Southwest
	Southwest transitional	Southwest transitional	Southwest transitional
	Kainuu + Savonian Wedge	Kainuu + Savonian Wedge	Central Finland + Savonian Wedge + South Kainuu
	East Savo	East Savo	East Savo
	Päijät-Häme	Päijät-Häme	North + Border Karelia + Ingria + Coastal
K=12	Southeast	Savo	North + Border Karelia + Ingria + Coastal
	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	Southeast Häme	Southeast Häme	Southeast Häme
	Far North	Far North	Far North
	Middle / North Ostrobothnia	South Karelia	Middle / North Ostrobothnia
	Southwest	Southwest	Southwest
	Southwest transitional	Southwest transitional	Southwest transitional
	Päijät-Häme	Päijät-Häme	East Savo
	North Karelia + North Savo	Central Karelia	Southeast Proper + Savitaipale / Lemi
	Kainuu + Savonian Wedge	Central Ostrobothnia proper	Kainuu + Savonian Wedge
	South Savo + Savonlinna transitional	Central Ostrobothnia highlands + North Ostrobothnia	West Savo
	West Häme	West Häme	West Häme
K=13	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	West Häme	West Häme	West Häme
	Far North	Far North	Far North
	Southwest	Southwest	Southwest
	Southwest transitional	Southwest transitional	Southwest transitional
	Central Finland	Middle / North Ostrobothnia	Middle / North Ostrobothnia
	East Savo	Southeast	East Savo
	Päijät-Häme	Päijät-Häme	North + Border Karelia + Ingria + Coastal

	Structure (diploid)	Structure (haploid)	K-medoids
	South Karelia	South Savo	South Karelia
	Central Karelia	North Karelia + North Savo	Central Karelia
	Southeast Häme	Southeast Häme	Southeast Häme
	Central Ostrobothnia proper	Kainuu + Savonian wedge + Ostrobothnia highlands	Kainuu + Savonian Wedge
	Central Ostrobothnia highlands + North Ostrobothnia	Border Karelia + Southeast Savo	West Savo
K=14	South Ostrobothnia	South Ostrobothnia	South Ostrobothnia
	West Häme	West Häme	West Häme
	Southeast Häme	Southeast Häme	Southeast Häme
	Far North	Far North	Far North
	Middle / North Ostrobothnia	Southeast	Middle / North Ostrobothnia
	Southwest	Southwest	Southwest
	Southwest transitional	Southwest transitional	Southwest transitional
	Päijät-Häme	Päijät-Häme	Päijät-Häme
	South Karelia	Savo-Vyborg transitional	South Karelia
	Central Karelia	Central Ostrobothnia highlands + North Ostrobothnia	Central Karelia
	North Karelia + North Savo	North Karelia + North Savo	East Savo
	Kainuu + Savonian Wedge	Central Ostrobothnia proper	Kainuu + Savonian Wedge
	South Savo + North Karelia + Border Karelia	South Savo + Savonlinna transitional	North + Border Karelia + Ingria + Coastal
	Border Karelia	Border Karelia	Central Finland

Table 2: Clusterings compared to one another. The cells shaded grey highlight the differences between the clusterings. On each row, the cells may be all white (indicating that the same cluster could be identified from all the clusterings), one cell may be grey and two white (indicating that two of the analyses agreed with one another while the third one disagreed), or one cell may be white and two have different shades of grey (indicating that the three analyses disagreed).

K	Str (diploid) vs. K-medoids	Str (diploid) vs. Str (haploid)	Str (haploid) vs. K-medoids
2	0	0	0
3	0,67	0	0,67
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0
8	0	0	0
9	0,33	0,33	0,33
10	0,30	0,30	0,20
11	0,45	0	0,45
12	0,33	0,42	0,50
13	0,31	0,46	0,38
14	0,21	0,29	0,50

Table 3: Percentage of disagreeing clusters with different analyses across different K values, calculated by dividing the number of disagreeing clusters with the K value (e.g. with K=3 two clusters out of three (67 percent of all the clusters) disagreed between the analyses.

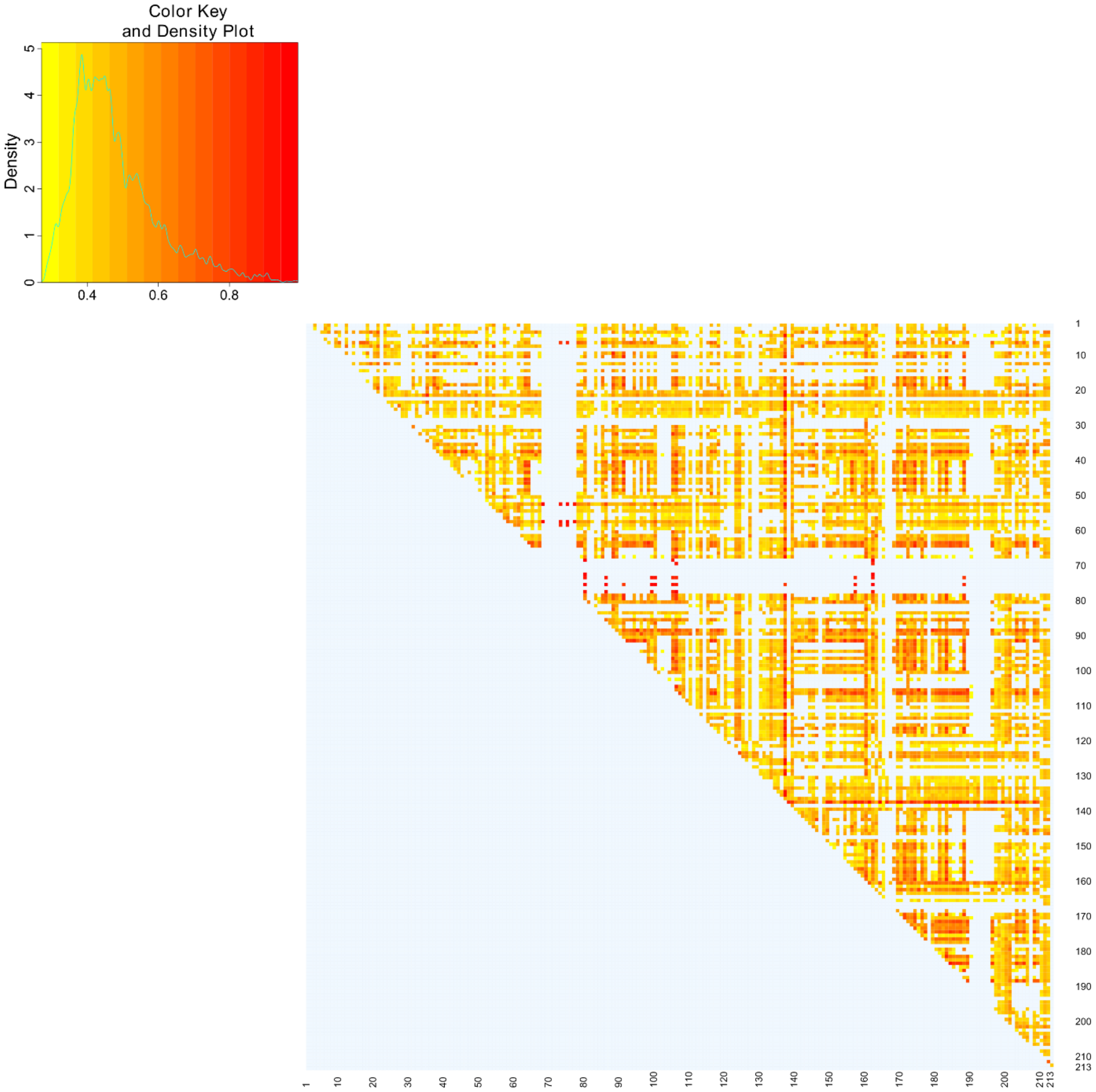

Figure 1: A linkage test heat map filtered by removing data points where potential linkage (Lp) value was less than a 25% of the highest Lp value in the results. This illustrates one way of identifying less reliable estimates.