UNIVERSITY OF TARTU
Faculty of Science and Technology
Institute of Computer Science
Computer Science Curriculum

Jose Rodrigo Flores Espinosa

Classification of human Y chromosome haplogroups based on dense and sparse genetic data using machine learning approaches

Master's Thesis (30 ECTS)

Supervisors: Kallol Roy, PhD
Monika Karmin, PhD

Tartu 2022

Classification of human Y chromosome haplogroups based on dense and sparse genetic data using machine learning approaches

Abstract

The genetic data of human Y chromosomes is classified into haplogroup categories based on the underlying phylogenetic tree, where a haplogroup represents a monophyletic clade on the tree. Current methods for the assignment of these categories work by representing a known human Y chromosome phylogeny as tree data structure. For an individual Y chromosome to be assigned a haplogroup using this representation, strategies based on breadth-first search (BFS) are often used. The tree is traversed in a manner that paths showing supporting evidence from mutations are further explored eventually leading to a leaf node and final classification. This strategy shows high efficiency when dense genotyping/sequencing data are available. However, in case of lower density genetic data such as genotyping arrays or ancient DNA data, BFS-based strategies often fail to reach a leaf node due to uncertainty and lack of information of where to go next.

In this work we leverage the increasing availability of world-wide panels of Y chromosome data with available curated haplogroup categories. We present a novel method on the application of a K-nearest neighbors classifier to both low-density and high-density types of data. The main goal is to assess the extent to which this approach can be useful in the challenging cases where BSF-based methods fail to produce a tractable and meaningful result. To achieve this, we have employed different DNA sequence encodings together with dimensionality reduction techniques. We have also investigated a novel method of DNA representation using Word2vec contextual embeddings. The DNA snippets are represented as text words and the whole DNA sequence is a text sentence. Encoding the DNA sequences in this manner gives rich contextual information that helps in haplogroup classification and can be extended to other applications in genomics.

The results show that classification accuracy is high (>98%) with next-generation sequencing (NGS) and genotyping arrays, high-density and lower-density data classes respectively. Performance however is low (<60% on average) when classifying ancient DNA data, which has the lowest level of resolution and higher levels of error. We observe that in many of the challenging cases KNN fails to correctly predict the label at its finest degree of resolution but does classifies correctly at the main category level which can be useful in practice.

Keywords: Y chromosome, Machine learning, haplogroup classification **CERCS:** B110. Bioinformatics, medical informatics, biomathematics, biometrics; P170. Computer science, numerical analysis, systems, control; B220. Genetics, cytogenetics

Masinõppe meetoditega inimese Y kromosoomi haplogruppide määramine tihedatest ja hõredatest geeniandmestikest

Lühikokkuvõte:

Inimese Y kromosoomi geeniandmeid klassifitseeritakse haplogruppide kategooriatesse vastavalt fülogeneetilisele puule. Monofüleetilisi klaade nimetatakse puul haplogruppideks. Nende kategooriate määramiseks esitavad praegused meetodid teadaolevat Y kromosoomi fülogeneesipuud puukujulise andmestruktuurina. Tihti kasutatkase üksiku Y kromosoomi haplogrupi klassifitseerimiseks laiutiotsingut. Puu käiakse läbi, uurides edsi vaid neid radu, millele on DNA-andmestikus olemas variandid, mille toel jõutakse leheni ehk lõpliku haplogrupi klassifikatsioonini. See strateegia on väga tõhus tihedate sekveneerimis- ja genotüpiseerimisandmestike puhul. Samas madalama tihedusega andmestike puhul – mõned genotüpiseerimiskiibid või vana DNA andmed - ei õnnestu laiutiotsinguga leheni jõuda ebamäärasuse tõttu, kuna pole piisavalt infot kuhu edasi minna.

Selles töös kasutasime ära järjest enam kättesaadavaid ülemaailmseid inimese Y kromosoomi täpsustatud haplogruppidega andmestikke. Rakendame töös k-lähimate naabrite (KNN) klassifikaatorit uudsel viisil erineva tihedustega andmestikele, ulatuses kõrgtihedatest sekveneerimisandmestikest kuni väga hõredate vana DNA andmestikeni. Töö peamine eesmärk on hinnata selle lähenemisviisi kasulikkust keerukate juhtude puhul, kus laiutiotsingul põhinevate meetodite abil ei õnnestu jõuda selgete sisuliste tulemusteni. Selle saavutamiseks kasutasime erinevaid DNA variantide kodeerimisi koos dimensioonide vähendamise tehnikatega. Me uurisime ka uudset meetodit DNA variantide esitamiseks, kasutades Word2vec kontekstuaalset vektoresitust. DNA variantide väljavõtteid esitatakse sõnadena tekstis ja kogu variantide genotüüpi lausena. Sellisel viisil kodeerimine lisab rikkaliku kontekstuaalset informatsiooni, mis aitab haplogruppide klassifitseerimisel ja seda võib rakendada ka muudele genoomika andmetele.

Tulemused näitavad, et klassifitseerimise täpsus on kõrge (>98%) uue põlvkonna sekveneerimise andmete puhul ja genotüpiseerimisandmete puhul, mis on vastavalt tihe ja hõre andmestik. Vana DNA puhul, mis on kõigw hõredam ja vigaderohkem andmestik, on aga täpsus oluliselt kehvem (<60%). Nägime, et paljude keerukate juhtude puhul KNN ei suuda õigesti ennustada täpseimat klassifikatsiooni, kuid suudab määrata põhikategooria ja see võib siiski olla praktikas kasutatav.

Võtmesõnad: Y kromosoom, masinõpe, haplogrupi klassifitseerimine
CERCS: B110. Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika; P170. Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria); B220. Geneetika, tsütogeneetika

Contents

1 Introduction

1.1 Human evolutionary genetics

The founding principle of evolutionary biology is that all living and extinct forms of life share a common ancestor and the genetic history of their development and evolution is contained withing their genomes (Fig. 1). In the case of *Homo sapiens*, we are able to extract and derive information about the demographic history of our own species by comparing, categorizing and analyzing the genetic sequences of living human individuals from worldwide populations in a systematic manner. This includes now ancient individuals whose DNA has preserved and also from closely related species such as other primates (M. Jobling et al. 2014). Knowing about our own biology and evolutionary past makes our present much more meaningful and interesting. This knowledge also has a multitude of ramifications and applications in the broader field of genomics and related disciplines.

1.1.1 DNA variation: From classical genetic markers to 'telomere-to-telomere' assemblies

The study of human molecular genetic variation has witnessed dramatic progress. The field has gone from the discovery and typing of small-scale insertion/deletion (INDEL) and single-nucleotide (SNPs) changes (Behar et al. 2008) to the completion of the sequence of the human reference sequence first draft in 2001; and all the way to continuous 'telomere-to-telomere' sequencing of complete chromosomes just a few years ago (Wrighton 2021). Many technological milestones have led to the development of the so-called next-generation and third generation sequencing technologies (Metzker 2010). This has triggered a true explosion also in the human genetic studies - genome-wide data is available from populations across the world. This sparked the developing of many methodologies for analysing sequenced genomes to address biological problems (Fig. 2). The amount of genetic data thanks to these advancements keeps increasing both quantitatively and qualitatively.

1.1.2 The human genome and its inheritance patterns

The human genome is diploid, much like in almost all other animals (M. Jobling et al. 2014). This means that within the nucleus of each somatic cell (those forming tissues) we have two copies of the genome. Each of these copies is about 3.2 billion nucleotide bases long (3.2 gigabases) and is divided into 23 physically independent portions called chromosomes that, based on homology, pair with their similar chromosome copy. In humans only 22 of them (the autosomes) are present in a diploid state, the last pair of chromosomes is different in males and females. Similarly to the other chromosomes

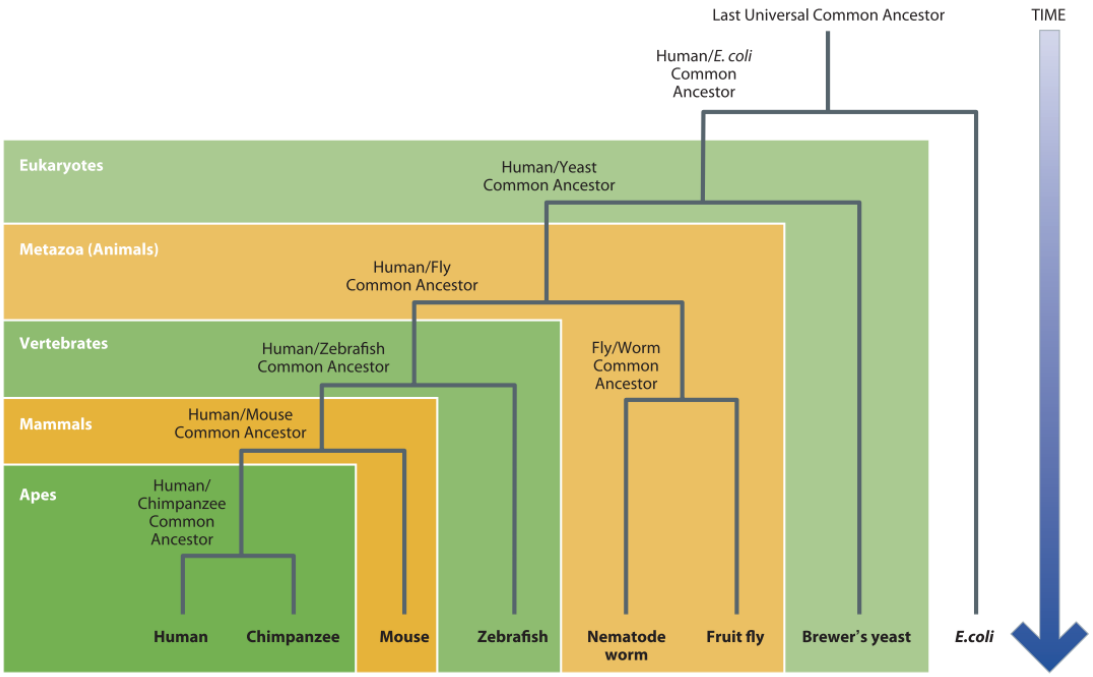

Figure 1. Simplified diagram illustrating the evolutionary relationships through time of various species including humans and how all derive from a theoretical last universal common ancestor (LUCA) organism. Figure re-published with permission of "Garland Science, Taylor Francis Group, LCC" and taken from the book "Human Evolutionary Genetics, 2nd Edition"; the permission was granted through Copyright Clearance Center, Inc.

females have two copies of X chromosomes in diploid state, while instead, males have a single copy of X chromosome and a single copy of Y chromosome, both in haploid state along most of their length. This last pair of chromosomes determines the sex of the individual. In addition to these 23 chromosomes, there is also the small mitochondrial (mtDNA) genome, which exists outside the nucleus of all cells in the form of a multi-copy short circular haploid genome in both sexes.

Not all the chromosomes show the same parent-children inheritance pattern. In a nutshell, haploid germline cells are produced by each parental sex in a process of mitosis followed by meiosis, where the genetic material is halved into a single set of haploid chromosomes. With the exception of mtDNA, Y chromosome and the X chromosome in male parents, as part of mitosis all the chromosomes undergo a process of recombination. This process sometimes referred to as "The great re-shuffler" is

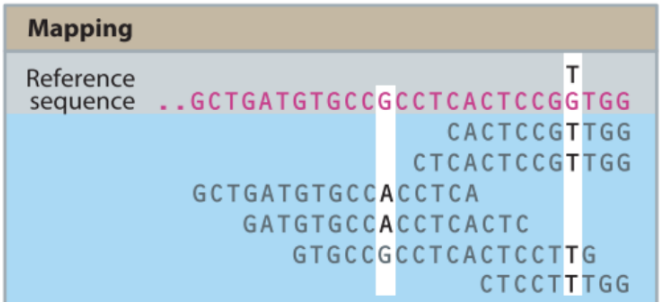

Figure 2. Simplified example illustrating the basic working principle of how we 're-sequence' genomes using modern next generation sequencing technologies. In this framework we first fragment the genome into small chunks that we can read. Then we align all these chunks to the reference and realize variation such as single nucleotide polymorphisms (SNPs) highlighted in the white columns. Figure re-published with minor modifications and permission of "Garland Science, Taylor Francis Group, LCC" and taken from the book "Human Evolutionary Genetics, 2nd Edition"; the permission was granted through Copyright Clearance Center, Inc.

responsible of generating new genetic variation in the resulting germline cells based on different possible combinations of the two DNA copies from their progenitors. During fertilization, the union of the male and female haploid sets of chromosomes result in a new complete diploid genome carrying information from both parents. Exceptions to this are mtDNA, X and Y chromosome. All paternal mtDNA is lost at the moment of sperm and ovarian fusion and this results in the the newly formed zygote inheriting mtDNA exclusively from the mother. On the other hand the Y chromosome passes from fathers to sons exclusively. The fusion of sperm carrying the Y chromosome (50% chances that occurs), results in the zygote becoming a male during development and this makes the Y chromosome effectively a male specific chromosome. The fact that the Y chromosome does not have a homologous copy prevents it from shuffling sequence during meiosis. The minor DNA differences eventually accumulate due to random mutations but not due to recombination and this gives the special properties for the study of human history and evolution (M. Jobling et al. 2014). Fig. 3 summarizes the general aspects of genome inheritance while Fig. 4 illustrates the specifics about recombining and non-recombining sections of the genome during this same process.

1.1.3 The advantages of Y chromosome as tool in human population genetics

Summarizing, the Y chromosome has three characteristics that make it specially suitable for evolutionary and population history studies:

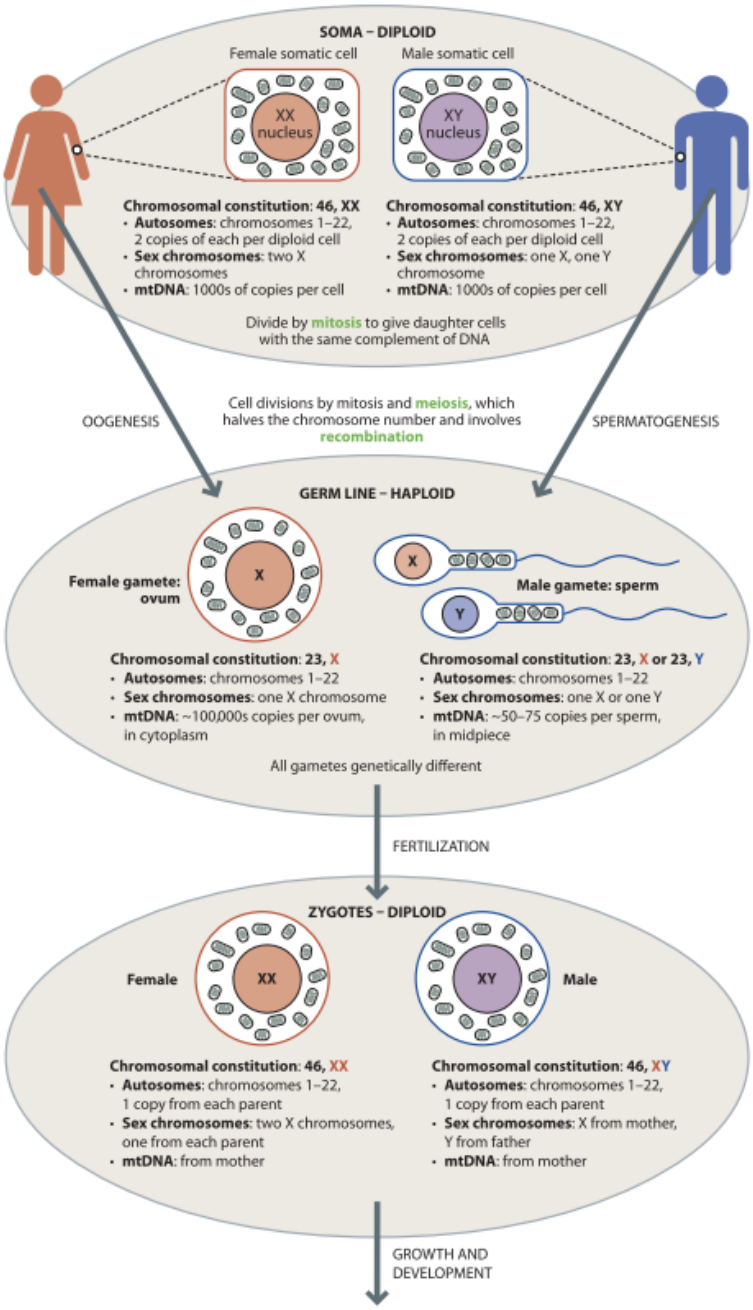

Figure 3. Overview of mitosis, meiosis, fertilization, the different types of DNA inheritance in the human genome and the portions under each of these categories. Figure re-published with permission of "Garland Science, Taylor Francis Group, LCC" and taken from the book "Human Evolutionary Genetics, 2nd Edition"; the permission was granted through Copyright Clearance Center, Inc.

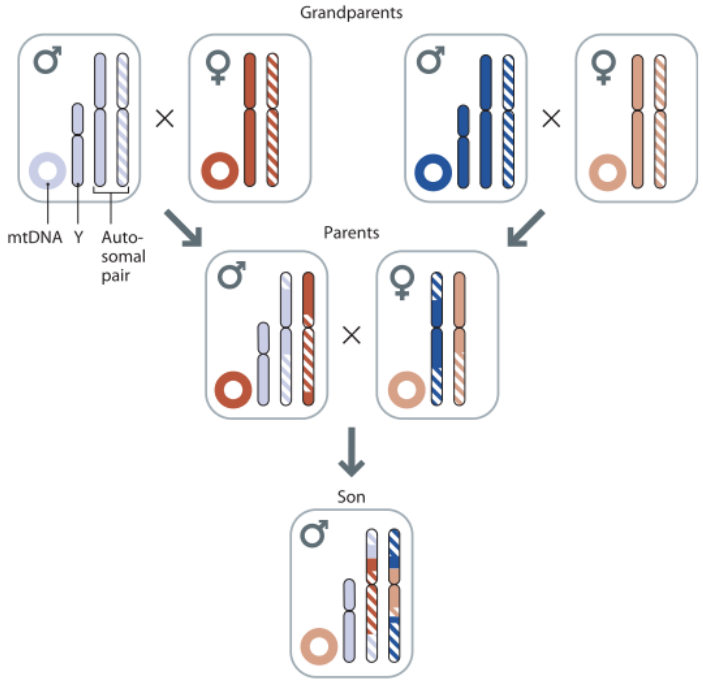

Figure 4. Condensed 3-generations schematic of the inheritance patterns for both recombining and non-recombining sections of the human genome. mtDNA and Y chromosome are inherited from a single parental line across generations; maternal in the case of mtDNA and grandfather-father-son in the case of Y chromosome and this particular example. The autosomes in contrast undergo recombination and do not have the same composition from one generation to the next. Figure re-published with permission of "Garland Science, Taylor Francis Group, LCC" and taken from the book "Human Evolutionary Genetics, 2nd Edition"; the permission was granted through Copyright Clearance Center, Inc.

1. Sex-specific inheritance - Y chromosome is inherited from father to son, so it has strict male-specific inheritance.

2. Haploidy. It comes in one copy/version as opposed to the autosomes which exist in two copies with some differences between them.

3. Lack of recombination. The absence of re-shuffling within the sequence at the moment of inheritance makes the copy being inherited almost identical to the original copy in the donor parent; any differences between them arise only thought the process of random mutation in the lapse of one generation.

These characteristics allow the construction of a phylogenetic tree that reflects the underlying true genealogical relationships among all males. From such tree it is possible to trace the evolution of paternal lineages within and among populations, their relatedness and sex-specif processes. The study of human history and evolution by means of Y chromosome phylogeography poses important challenges too, particularly in light of more recent computational and analytical techniques that allow to answer similar questions of population history but using information from the entire rest of the genome (autosomes) (Lazaridis et al. 2014). Importantly, the use of haploid genetic material like the Y chromosome remains an important piece of evidence in all studies aiming to reconstruct our recent and ancient history (Underhill and Kivisild 2007; Underhill and Kivisild 2007; Karmin et al. 2015; G David Poznik et al. 2016).

In addition to evolutionary genetics, the study of Y chromosome is quite relevant to genealogical research, forensics and medical conditions linked to the Y chromosome such as male infertility (Hallast et al. 2021).

1.2 Y chromosome haplogroup classification

1.2.1 Y haplogroup nomenclature systems

After two decades of steady progress on the discovery and study of various forms of genetic variation on the Y chromosome (Casanova et al. 1985; M. A. Jobling, Pandya, and Tyler-Smith 1997), the beginning of the new millennium saw new techniques being developed and the first comprehensive human Y phylogenies based on different human populations started to emerge. At the beginning each scientific group developed its own way to categorize the diversity they started to observe (Capelli et al. 2006). In 2002 however, a year before the reference sequence of the Y chromosome was published and release for public use, a group of researches in the field formed the Y Chromosome Consortium (YCC) and set to standardize the various existing nomenclature systems that were being used in the literature to define haplogroups (Hammer 2002). This resulted in the first Y chromosomal tree with world-wide representation; it was build based on 245

markers, most of them binary SNPs, and represented a set of 153 haplogroups/branches with standard nomenclatures to which markers and their state (ancestral or derived) could be mapped to for any individual in order to know its position on the tree. According to this nomenclature main haplogroups are given a letter from the alphabet and subclades are named with alternating alphanumeric labels starting with a digit (e.g. N1a1b).

In 2005 the https://isogg.org/ was born and adopted the YCC nomenclature system. Although its adoption has not been universal and other nomenclature systems addressing some of its disadvantages have been proposed (Karmin et al. 2015), the YCC nomenclature remains the most widely used naming system for Y chromosome haplogroups. For already many years up until now ISOGG continues to be the organization that maintains and curates the lists of mutations and the tree phylogeny stemming from them. Nowadays almost any set of markers typed on the Y chromosome for a given individual can be reliably mapped to the YCC/ISOGG nomenclature system resulting in the classification of this individual into a particular haplogroup.

1.2.2 Early automated approaches to Y chromosome classification

Y chromosome haplogroup classification remained a highly manual process and the only reported automatized method in the literature was a proprietary algorithm developed by the mid 2000's whose input consisted of Y-STR markers (a type of genetic variation polymorphic for the number of small repeated motives of DNA in each individual/haplogroup) and provided a prediction based on a compiled private database. In 2005, the first web-based and fully described method was published and it was based on a public compilation of Y-STR allele (one of two or more versions of DNA sequence (a single base or a segment of bases) at a given genomic location) frequencies for each haplogroup and a 'goodness of fit' approach (Athey 2006). This work does not aim provide a comprehensive revision of all methods and tools developed for this purpose since then but rather aims to provide a short summary. From the publication of (Athey 2006) and up until the end of 2010 only some variations and extensions to this method were developed. Relevant for this work, in 2008 the only work in the literature of Y haplogroup classification using machine learning (ML) approaches was published (Schlecht et al. 2008). The approaches described here however used Y-STR variation as input data (the only cost-efficient data retrievable at the time) and offer no insights into their usability for sparse genetic data. In the following sections the terms sparse and low-density genetic data are used interchangeably.

1.2.3 Modern automated Y chromosome classification strategies and limitations

It was only until 2014 and then 2016-2018, with next generation sequencing technologies and large sequencing projects becoming ubiquitous, that the first automated methods for Y chromosome classification based on larger and growing Y-SNP variation became available and end-users with no prior knowledge of phylogenetic reconstruction or Y chromosome phylogeny could feed their high-throughput genetic data to an algorithm and get back a haplogroup label.

A handful of recent methods have also been published with a focus on more specific cases of Y chromosome classification. One of these methods is presented in (Severson et al. 2018) and it exploits and uses high-density array-based genotyping data in particular as opposed to high-density data based on NGS or sparse genetic data such as aDNA or low-resolution genotyping arrays. Another example is presented in (Martiniano et al. 2020a), the only available method to date with a focus on classifying aDNA into haplogroups by placing new individuals into the Y chromosome phylogeny.

In this work we aim to provide and test new additional methodological frameworks for Y chromosome classification that are suitable for any type of SNP-based genetic data but particularly suitable for the challenges posed by sparse genetic data (aDNA, targeted sequencing, low-density genotyping arrays). For this purpose we present results based on machine learning approaches as well as a variation of the more standard tree-based search approach to Y chromosome prediction.

Current methodologies to classify human Y chromosomes into haplogroups often rely on representing the known Y phylogeny and their constituent mutations as a tree data structure (Martiniano et al. 2020b; G David Poznik 2016a; Jostins et al. 2014; Severson et al. 2018; Ralf et al. 2018). This representation is then traversed using variations of breadth-first search (BFS) in which the new Y chromosome under classification is interrogated at each step of the traversal in order to give direction to the search, which generally avoids extending the search in branches found with no supporting evidence of continuation. The moment one of the tips of the internal tree representation is reached, a label corresponding to that node gets assigned to the new sample. There is a caveat in this approach though: it works generally well with high-quality and high density genotyping data but often fails to even produce any meaningful or approximate label in cases of highly sparse genetic data such as ancient DNA (aDNA), low-density arrays or cases where the set of mutations typed in a sample do not overlap well enough with the set of mutations in the particular tree representation that the method uses. The main reason why BFS-based approaches fail to produce a label in these cases is mainly due to the lack of information (genotypes) at certain nodes within the tree and the impossibility (or lack of appropriate heuristic) on how to proceed with the search See Fig. 3. A non-naive

(exhaustive traversal of the tree in all branch directions) solution to this problematic is not trivial as one can tell from the lack of methods available. Only one recent method addresses the problem to our knowledge Martiniano et al. 2020b.

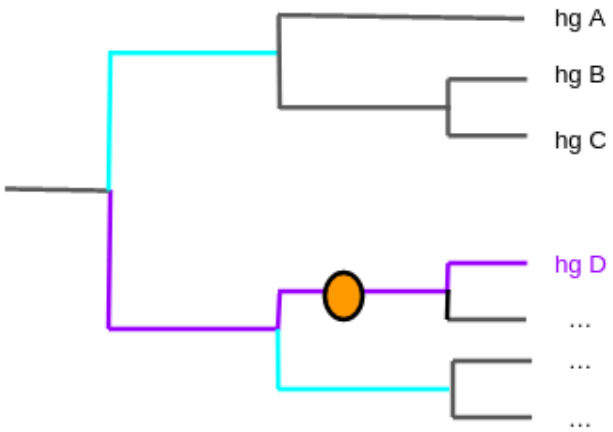

Figure 5. Cartoon representation of a Y chromosome phylogeny section that illustrates the problem often encountered by breath-first search -like approaches in the face of sparse genetic data. The line in magenta represents the path of nodes and mutations leading to the true haplogroup (hg D) of a hypothetical individual being interrogated. The circle in yellow represents the standstill point at which the algorithm cannot continue further down the path due to a lack of information on that particular branch. Cyan lines represent paths in the que that have been checked but left since there is evidence for not to continue in those directions.

1.3 Machine Learning techniques for Y chromosome classification

Currently there are no available machine learning methods that exploit the large and growing number of high-quality labelled data on Y chromosomes that is publicly available in order to produce models for haplogroup prediction based on current forms of dense and sparse genomic data. In this work, we have explored K-nearest neighbors (KNN) as a classifier for the assignment of haplogroups (classes) to Y chromosome genetic data. Additionally we present an attempt to model genetic DNA using contextual word-based embeddings. To do this nucleotide 'words' are formed based sub-strings of fixed size within the Y chromosome and their corresponding embeddings (real-valued vectors) are estimated using a shallow neural network that maps genetic words with similar adjacent context to similar embeddings. We think this can be used not only as an alternative data

representation to ML models but also potentially useful in other applications like an ML-based generator of DNA sequence that is indistinguishable from real DNA.

1.3.1 K-nearest neighbors classifier

In statistical non-parametric unsupervised learning, the K-nearest neighbors algorithm is a method for both classification and regression tasks (Casella, Fienberg, and Olkin 2017). In the case of classification, this algorithm predicts the class of new data points using the vote of close data points (neighbors in a multi-dimensional space of features) where each neighbor votes for its own class and the most common class among the k-closest neighbors becomes the predicted label. The optimal value of k is data-specific and has to be defined *a priori*. Typically various values for k are evaluated and possible ties in the voting are resolved at random if the chosen k value is an even number. In order to estimate closeness, or distance between neighboring data points, different metrics of distance can be used. A common metric when working with continuous variables is Euclidean distance, while other metrics such as Hamming distance are also used for discrete variables.

For both classification and regression a common approach is to use weights so that the votes of closer neighbors get a higher say in the final decision compared to more distant data points which still are considered as neighbors. A common weighing setup consist in giving each neighbor a weight of $1/d$ where d is the distance to the neighbor. Attention has to be taken when choosing the features and their scaling, since non-relevant features or scale units not corresponding with their importance result in degraded performance. Since often each feature is expressed in different scales, it is common to standardize the features in order to have mean equal to 0 and variance equal to 1 Casella, Fienberg, and Olkin 2017.

When the number of features in the data is very large and/or suspected to be redundant, the number of dimensions is often reduced to a smaller set of 'most relevant' features using extraction methods and other types of dimensionality reduction techniques such as Principal Components Analysis (PCA), t-SNE embeddings, Isomap, etc. This reduced data representation can then be used as input for KNN. Regarding model training, the set of neighbors can be thought as the training set although there is not exactly a training step in this algorithm but rather a storage of the training data by means of a fast and efficient indexing data structure such as Ball or KD Trees) (Barupal and Fiehn 2019).

1.3.2 Principal component analysis (PCA)

It is nowadays typical to work with datasets that have many features and are highly dimensional. This situation is often referred to as the 'curse of dimensionality' and

examples of data exhibiting such property span a wide range of areas including the field of medical and human genetics. PCA has been widely used in the field of genetics for various decades now, mostly for the purposes of data exploration and study design (Reich, Price, and Patterson 2008). In this work, we used PCA to explore the impact that reducing the number of dimensions in the data set has on the efficiency of KNN for classification.

Principal component analysis (PCA) is a multivariate analysis technique (unsupervised learning) that reduces the number of dimensions in a given data set while reducing as much as possible information loss. It does this by first computing the principal components (PCs) of the data and then using them to perform a change of basis into a dimensionally reduced representation (often only a few PCs) of the data that retains as much as possible of the original variance. The PCs - one for each dimension - have the property of being unit vectors that correspond to the direction of the best-fitting line which is that one minimizing the average squared distance from the points to the line or equivalently, the line maximizing the explained variance along each of the dimensions. Each $i-th$ PC is uncorrelated and orthogonal to the first $i-1$ PCs. The PCs correspond to the eigenvectors of the covariance matrix of the data and computing them using either eigendecomposition of the covariance matrix or singular value decomposition of the data matrix is often computationally preferred.

1.3.3 Word embeddings

Natural language processing (NLP) is a discipline at the intersection of statistical and machine learning, computer science and linguistics. A core goal of NLP is to develop computational frameworks capable of representing and fully understanding text and voice data as well as to produce the same type of representations as output in ways and speed that humans do. The applications of such frameworks and technologies are wide and span a wide variety of applications. In order to achieve this it is necessary to represent the basic units of language (words) in a manner that computers can process it efficiently and that also captures desired properties of language. There are multiple techniques that focus on different aspects of language such as syntax, semantics, etc. and aim to to build such mapping/representation. One of them is word embeddings.

Word embeddings is a form of text representation that places a central role in the context (neighboring words) in which words are used in language as a conveyor of meaning. Floating point vectors representing these words have the resulting property of being close in meaning to other numerical vectors encoding words with similar meaning, making this form of language representation particularly suitable for applications where context is important. In addition, embeddings capture relationship between words with a level of detail that is adjustable via the size of the embedding for each word; larger

embeddings generally being able to capture finer levels of relation and meaning among words.

The creation of such mapping involve a transformation from a higher dimensional space in which words and their context are represented, to a lower dimensional space or vector representing each word. There are various techniques in NLP to create word embeddings. The most promising and used ones are Term Frequency-Inverse Document Frequency (TF-IDF); produces a statistical measure corresponding to the relevance of words in a text, Word2Vec (Mikolov et al. 2013); learns associations among words that capture semantic meaning, Global Vectors (GloVe) (Brennan et al. 2017); extends the framework of Word2Vec with the goal of capturing global context information as opposed to local context (Word2Vec), and Bidirectional Encoder Representations from Transformers (BERT); a transformer model architecture developed on the basis of so-called attention-based learning and aims to capture fine-level semantic associations among words. In this work we use a shallow neural network implemented using a Keras Embedding layer that takes an integer-based representation of a matrix of nucleotide words of a given size and outputs one embedding vector representation for each of these words (See Methods section 3.7.3).

2 Objective

Even though the study of human history based on genetics has considerably shifted towards genome-wide analysis in the last years G David Poznik et al. 2016, Y chromosome haplogroup classification remains important and relevant not only in informing studies on human history and evolution but also in several others areas of investigation such as forensics and medical genetics Hallast et al. 2021. Most current methodologies to classify human Y chromosomes into haplogroups (Martiniano et al. 2020b; G David Poznik 2016a; Jostins et al. 2014; Severson et al. 2018; Ralf et al. 2018) build first a tree representation of the Y known phylogeny and then traverse it by interrogating the Y chromosome data under classification with the aim of reaching a tip corresponding to the label. However, this strategy poses practical challenges when the data interrogated is sparse, noisy and/or do not overlap well enough the specific set of mutations used by the specific implementations Severson et al. 2018. Often this results in no labeling being produced and the need to manually check the data to produce a haplogroup label. In this work we wanted to explore machine learning approaches that could leverage the increasing availability of high-quality and manually curated data in order to perform haplogroup classification in a manner that did not require building a knowledge of an *a priori* tree representation. In order to achieve this we explore the possibilities of K-nearest neighbors as a classifier based genetic data at various degrees of quality and information density. In addition, we also explore the suitability of representing

genetic data as word-like context-sensitive embeddings and their use as input for our classification models.

3 Methods

3.1 Pre-processing of high-density data

3.1.1 Sources of genetic data

High-density genotyping data in the form of FASTQ (Illumina technology) and varFile (Complete Genomics technology) formatted files were obtained and downloaded from various public sources including 1000 Genomes Project (1000K; Byrska-Bishop et al. 2021), Human Genome Diversity Panel (HGDP; Bergström et al. 2020), Estonian Genome Diversity Panel (EGDP; Karmin et al. 2015), Simons Genome Diversity Panel (SGDP; Mallick et al. 2016), Siberian and Northwestern European Panel (Siberian NEE; Wong et al. 2017) and several other studies whose data has been published and made available to the public (See Apendix section II for complete details). With the exception of targeted sequencing data obtained from the sequencing service bigY from GenebyGene, all data are derived from whole-genome sequencing (WGS). In all cases the coverage of the data corresponds to standard high-quality 30x or more when estimated at the genome level. A unit of 1x coverage corresponds to the average number of times that any given base of the genome has been sequenced and covered 1 time by a short read. Altogether these data represent 3079 Y chromosomes from male individuals around the world at various degrees of haplogroup representation (Table 1).

3.1.2 Read mapping

Read mapping was done differently for Illumina and Complete Genomics data. Illumina short-read sequencing data were retrieved in the form of FASTQ files and aligned to the human reference genome version hs37d5 using the Burrows-Wheeler aligner (BWA v7.12) which resulted in Binary Alignment Map (BAM) files as output. Reference hs37d5 contains a 59 Mb long Y chromosome sequence where the both PAR recombining regions have been N-padded. After alignment, exact duplicated reads mapping to the same exact location of the genome were detected and one copy of them marked as duplicate using Picard (v2.25). This in order for these reads not to be considered during variant calling since they are likely PCR artifacts. All reads were locally re-aligned around INDELs and base qualities were re-calibrated using The Genome Analysis Toolkit (GATK v3.8) tools RealignerTargetCreator/IndelRealigner and BaseRecalibrator/PrintReads respectively. All intermediate processing of BAM files was handled with SAMtools (v1.12). This included the final extraction of the mapped reads corresponding to the Y chromosome.

Complete Genomics data did not require the mapping processing step since this now legacy commercial sequencing service performed the mapping using proprietary software and provided varFiles containing genetic variants already called . Data corresponding to the Y chromosome was extracted from varFiles using UNIX and AWK commands.

Table 1. High-density genotyping datasets

Total Haplogroups	Publication / Sequencing Type	Y chromosomes
A 58 B 26 C 124 D 32 E 541 F 6 G 84 H 95 I 151 J 281 K 8 L 63 M 15 N 214 O 374 P 1 Q 146 R 833 S 7 T 20	***1000 Genomes Project (1000G)*** *(Byrska-Bishop et. al 2021) / Illumina WGS*	1233
	Human Genome Diversity Panel (HGDP) *(Bergstrom et al. 2020) / Illumina WGS*	536
	Estonian Genome Diversity Panel (EGDP) *(Karmin et. al 2015) / Illumina WGS*	303
	Complete Genomics Compilation (CGC) *(Various projects and publications) / Complete Genomics WGS*	376
	Simons Genome Diversity Panel (SGDP) *(Mallick et. al 2016) / Illumina WGS*	175
	Siberian NEE Panel (SNEEP) *(Mallick et. al 2016) / Illumina WGS*	16
	GenebyGene BigY Compilation (GBC) *(Various projects and publications) / Illumina (targeted)*	440
	Total	3079

3.1.3 Variant calling

Illumina-based BAM files containing the mapped reads on the Y chromosome were used as input to GATK (v3.8) HaplotypeCaller using parameters ploidy 1, min-base-quality-score equal to 20 and emitRefConfidence BP_RESOLUTION. The latter parameter was used in order to get genotype information of every single base in the Y chromosome.

This resulted in one Genome Variant Call Format (gVCF) file per chromosome, which were then processed with GATK tools CombineGVCFs and GenotypeGVCFs in order to produce a single multi-sample Variant Call Format (VCF) containing genotypes for all positions across all Y chromosomes in the data set.

Complete Genomics data did not require variant calling, since the available varFiles already contained non-variable and variable sites. Data in varFiles format were converted in to VCF files with genotypes for all positions in the chromosome using the cgivar2gvcf tool.

3.1.4 Quality filtering

Quality filtering of genotypes from Illumina data was performed using BCFtools (v1.15). The minimum amount of supporting reads for a variant genotype to be retained was set to 2 and 4 for the cases of GenebyGene targeted sequencing and WGS data respectively, while the minimum quality for invariant reference genotypes was set to 30 (log Phred-scale). Cases were these two criteria were not met were switched to missing state. All variable sites from Complete Genomics data already in VCF were switched to missing state if any of the low-quality tags ("lowqual", "ambiguous") specific to this platform was present.

Data from both technology platforms was kept in two versions for downstream analysis: masked and unmasked. Masked data included positions present at the PHIP whereas unmasked data contains genotypes for all sites in the Y chromosome, both invariant and variable.

3.1.5 Haplogroup labeling

Due to the high-quality and high genotyping density of both WGS and high-coverage targeted sequencing in this data set, haplogroup labeling was done primarily using the tree-traversing algorithm yHaplo (G. David Poznik 2016b), which is well suited for this type of data. The labels assigned by this method are based on ISOGG, which is the most common Y haplogroup nomenclature system and thus suits well the purpose of making downstream comparisons between these labels and manually derived haplogroups based on more difficult to classify data (e.g aDNA, low-density arrays) and which are also available as ISOGG labels.

3.2 Pre-processing of low-density data

3.2.1 Sources of genetic data

Low-density genotyped Y chromosome data in the form of EIGENSTRAT formatted files were obtained and downloaded from public sources Allen Ancient DNA Resource

(AADR) and the Affymetrix Human Origins data set. The AADR is one of the biggest compilations of public genetic data (from genotyping arrays and NGS but in both cases low-density final genotyping) from a multitude of aDNA studies carried on during the last decade. The Affymetrix Human Origins data contains individuals overlapping the Human Genome Diversity Panel but have been included since they by far do not represent duplicated instances of the same sample; array data has its own error distribution and represents an arbitrary subspace of the variation assayed using high-density WGS. Array data thus models better the errors and positions found in genotyping arrays, therefore once combined with its high-density counterpart we expect to to increase classification efficiency overall. Altogether these data represent 6,486 Y chromosomes from male individuals around the world (and time) at various degrees of haplogroup representation (Table 2).

Table 2. Low-density genotyping datasets

Total Haplogroups	**Publication / Sequencing Type**	**Y chromosomes**
A 16 B 17 C 148 D 23 E 207 F 5 G 229 H 81 I 530 J 272 K 6 L 58 M 17 N 111 O 135 P 5 Q 422 R 1080 S 14 T 31	***Ancient DNA Harvard Compilation (v50)*** *(Various publications) / low coverage*	1624
	Ancient DNA Harvard Compilation (v50) *(Various publications) / very low coverage*	1232
	(Lazaridis et al. Nature 2014) */ array data*	551
	Total	3407

3.2.2 Format transformations

Eigenstrat files were converted to binary PLINK format using the tool convertf from Eigensoft and from this format to VCF using a combination of PLINK (v2.0) commands and BCFtools (v1.15). Special attention was taken in order to ensure that this conversion process recorded in the final VCF the reference allele as the nucleotide that is actually present in the reference human reference (hs37d5), since PLINK is a reference-free format and a regular conversion procedure chooses the reference allele based on which is the most frequently observed in the cohort. The purpose of these steps was to have the same the format (VCF) across all data.

3.2.3 Quality filtering

In contrast to high-density WGS data, sparse genetic data sets were filtered at the individual level in order to discard Y chromosomes with extremely low coverage (<0.1x genome-wide), not meeting standards of quality (Assessment status != 'Pass') or not having a good enough level of resolution in their haplogroup labeling (e.g. haplogroup 'HIJK'). Low-density genetic data and in particular aDNA data varies widely in terms of final coverage and degree of sparsity. All Y chromosome passing the filtering criteria were divided in two coverage categories based on genome-wide estimates: 'low'; more or equal than 1x, (estimated mean of 3.03x, median of 1.19x and maximum of 56x), and 'very low'; between 0.1x and no more than 1x. The expectation was that the 'very low' bracket would be more difficult to correctly classify than the 'low' bracket.

3.2.4 Haplogroup labeling

Haplogroup labels based on ISOGG nomenclature were already available for the AADR public compilation. It is important to note that for most of these data, haplogroup labels have a certain level of manual curation not only by the authors of each specific study that originally generated the data but also from the many researchers and citizen science initiatives interested in genealogy and aDNA studies. To our knowledge there is no better curated set of Y chromosomes with already available haplogroup labels with some level of manual curating than this one. In the case of the Affymetrix Human Origins data set. we had available the corresponding labels based on their higher quality counterparts from the high-density data sets so these were matched and assigned See Apendix section II for further details.

3.3 Panel of Y haplogroup informative markers

A list of curated haplogroup-informative (shared among individuals and not private variation) positions on the Y chromosome was assembled into a "Panel of haplogroup-informative positions" (PHIP) by taking the union of three already curated position

panels: Ralf et al. 2018, G. David Poznik 2016band Severson et al. 2018. The first two lists of positions are derived from the ISOGG database (2016 and 2019 respectively) and the third is enriched for positions included in the Illumina Multi-Ethnic Global genotyping array (Severson et al. 2018). Additionally, all positions included in both the Allen Ancient DNA Resource (AADR) and the Affymetrix Human Origins array were added to the union. Although genotypes on positions outside the genotyping arrays would be always missing state in Y chromosomes only genotyped in arrays, this ensured that for Y chromosomes genotyped using NGS, we always have the precise genotype (either non-variable or variable) at all positions present in arrays. In total this procedure resulted in 65,177 positions relevant to Y chromosome haplogroup classification.

3.4 Final merging, filtering and masking

Data from both the high and low density panels were merged and subset to the 65,177 positions in the Panel of haplogroup-informative positions (PHIP) using BCFtools (v1.15). In this process all variants of length or variant representation > 1 in the VCF format were discarded and only single-nucleodide variants (SNVs) were retained. This resulted in a final total cohort of 6,486 individuals and 65,177 positions.

3.5 Label shortening

ISOGG-based Y haplogroup classification consists of categories that grow in length the more refined the subcategory becomes. The general simplified rule in the nomenclature system is that a given category always branches first into a digit and then into a letter, and so on in an alternating manner (e.g. R1a1b1b1b1b2a). For each haplogroup label in the data set we produced two alternative corresponding labels with a reduced level of refinement following these rules. "root label" (root-label or r-label) consisted simply on the most general category (first letter) of the haplogroup. "2-degree shortened" (2d-label) excluded two pairs of digit/letter from the end of the label. Exceptions to this were labels with a length equal to 6 and shorter or equal than 5. In the first case only one pair of digit/letter was removed from the end while in the second case the label remained as such. These criteria are arbitrary.

3.6 Genome data representation

The Variant Call Format (VCF) offers great advantages and availability of powerful tools in order to handle the data and perform the various transformation and merging operations we needed to do on it. However, it has a small caveat for some of the analysis we wanted to do. PLINK format stores the actual nucleotide ('A', 'C', 'G', 'T', 'N') present at any given sample at any given position. In contrast, VCF stores an index (0,1, .., N) that corresponds to 0 if the observed genotype is the same as in the reference genome,

or N digit, where N corresponds to the N-th genotype that has been observed in the VCF cohort of individuals at that same position. While this type of variant encoding is great for reducing the size of the VCF in both its text-based form and binary compressed (BCF) version, it is thus required to perform lookup operations on the listed genotypes in order to realize the exact meaning of the index at each position/row in the format. Models and data representations used here benefit from 1) data representation in the form of nucleotides without the need to perform lookup operations and 2) still a lower level of data representation in the form of numbers. For this reason we encoded the data in two versions in addition to the VCF-derived format and used the most convenient encoding (or transformation between them) depending on the exact model used or its implementation. These data transformations were done using Python programming (v3.7.13) and libraries Pandas (v1.3.5) and Numpy (v1.21.6). This resulted in the following three matrices/encodings which are also illustrated in Fig. 6:

1. 'Reference-based (RB) Representation'. Genotypes are encoded as 0, 1, 2, 3 or '.' and the meaning of each integer/character is dependent on the genotype of the reference at each particular row. 0 in the first row can be equivalent to genotype 'A' but 'T' in another given row/position. The character '.' denotes missing data.

2. 'Nucleotide-based (NB) representation'. Genotypes are encoded as the exact nucleotide 'A', 'C', 'G', 'T' or '.' (missing data) seen at each position/sample (row/column). This representation is generated based on the RB data and lookup operations to the reference and possible alternate genotypes at each position across the data samples (Y chromosomes).

3. 'Integer-based (IB) representation'. Genotypes are encoded as integers 1, 2, 3, 4 or 5. Unlike the RB representation, here the following mapping nucleotide to integer is always valid: 'A':1, 'C':2, 'G': 3, 'T':4, '.':5. This representation is build from the NB data.

3.7 KNN implementation for Y haplogroup classification

K-nearest neighbors (KNN) as a supervised model for classification was implemented using the Python library for machine learning scikit-learn Barupal and Fiehn 2019. In all cases where we fit a KNN model we first used GridSearchCV() to make a search for the best combination of hyper-parameters given the specific input data. This search included parameters 'n_neighbors' : [1, 3, 5, 7], 'weights' : ['uniform', 'distance'], 'metric' : 'Minkowski', 'p' : [Manhattan','Euclidean'], 'algorithm' : ['Ball_tree', 'KD_tree', 'brute'], 'leaf_size' : [10, 20, 30], 'cross-validation' : [5] and 'n_jobs' : 20. All hyper parameter grid searches were run using computational resources from the High Performance Computing Center at Tartu University.

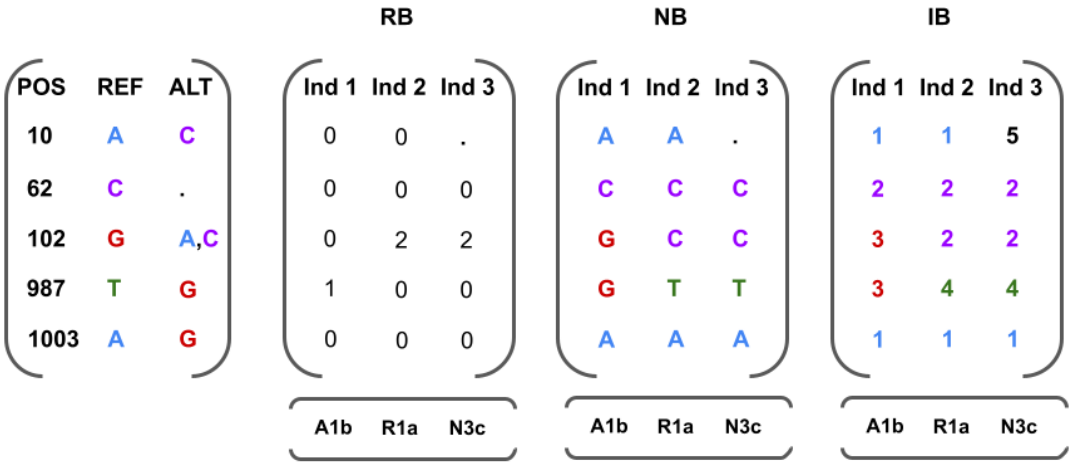

Figure 6. Data and intermediate data representations used as input for the ML models. RB, NB and IN correspond to Reference-based, Nucleotide-based and Integer-based data representations. 'POS', 'REF' and 'ALT' stand for position, reference and alternate alleles/genotypes, respectively. Each column represents an individual with its hypothetical haplogroup category shown at the bottom.

3.7.1 Data partitioning

The complete data set consisted of 6,486 samples. Of these, 3,079 and 3,407 are high-density and low-density respectively (Table 2 and Table 1. Within the low-density data, 551 are from genotyping arrays, 1,624 are 'low' (>=1x) aDNA category and 1,232 are 'very low' (>0.1x and <1.0x). From the beginning of the study a 'novel' partition was created by randomly sub-sampling 5% of samples from each of these subcategories. This partition was never used again but for the final validation of the best performing model. The 6,165 (95%) rest of the data was used for regular splits between training and test partitions. Training samples always consisted of 80% of samples from each of the data classes ('high', 'array', 'low' and 'very low'). Testing samples on the other hand included the rest 20% from each data class.

3.7.2 Standardization and PCA

Standardization of the data was performed using StandardScaler() from scikit-learn Barupal and Fiehn 2019. This transformation involves centering and scaling of features in the form of $z = (x - u)/s$, where u and s are the mean and standard deviation on the training samples. u and s values calculated on the training samples were used in the standardization of the test samples. PCA was performed using the library PCA

from scikit-learn Barupal and Fiehn 2019. This implementation uses Singular Value Decomposition (SVD) to find eigenvectors.

3.7.3 Word embeddings based on genomic data

Word-based embeddings are estimated using the nucleotide-based representation of the data. To form 'words' out of individual nucleotides we truncated the initial data matrix of 65,177 columns according to the word size we wanted to use. For word sizes 1, 2 and 10 data was truncated to 65,170 columns while for size of 100 we used 65,100 positions. All newly formed 'words' were uniquely mapped to integers starting from 1 in the order they appeared in the data. For example, using a word size of 10, the first word 'ACTGCTTAGC' would be mapped to 1, the second different 'AGT.CTAGCT' would be mapped to 2, etc. These integers were then used in the data matrix to represent the nucleotide-based words. The number of columns always remains constant for any given word size and there is no need of extra padding at the end of any row. We used these data as input to a neural network in order to estimate the embeddings for each word. To build the model we used the Embedding layer of TensorFlow/Keras, which receives vectors of integers as input and returns for each word/integer a unique real-valued vector of size N as its corresponding embedding. We chose N depending on the word size. Words of sizes 1, 2, 10 and 100 were represented as embeddings of sizes 1,1,2 and 10 respectively, thus implying a reduction on dimensionality in the final data matrix with the exception of word size 1. Finally each integer in the data matrix was replaced with its embedding word representation and the data was used as input for the KNN classifier. See Apendix section II for complete details.

4 Results and Discussion

4.1 Reference-based vs Nucleotide-based DNA representation

The standard format to represent DNA variation across multiple individuals and positions in the genome encodes genetic variants as an integers whose nucleotide(s) actual sequence depends on the observed nucleotide on the reference genome at that same location (Fig. 6). While this form of DNA representation offers great compression advantages by turning most of the data into 0 and 1 integers, the loss of explicit information of the exact nucleotide state (A,C,G,T) should correspond to a decrease in efficiency for some machine learning algorithms such as KNN. We tested this hypothesis and measured the differences in resulting accuracy. Nucleotide-based representation was derived from the reference-based data. Since KNN requires numerical data to compute distances, each nucleotide was mapped back to an integer (integer-based representation) using a unique mapping across the data so that each integer always encodes the same nucleotide (See

Methods section 3.6). We partitioned the data into training and testing samples following what we referred to as partition 'schema 1' (See Methods section 3.7.1). This partition strategy consisted on making training sets that contained 80% of samples from each of the different data classes into consideration ('high', 'array', 'low' and 'very low') and testing sets with the rest 20% of samples from the same data classes. The best hyperparameters under this model was ['algorithm': 'kd_tree', 'leaf_size': 10, 'n_neighbors': 1, 'p': 2, 'weights': 'distance']

Table 3. Reference-based vs Nucleotide-based representations

Data class	Data rep.	Acc.	Pr.	Rec.	F1
'high'	NB	0.89	0.87	0.89	0.87
	RB	0.80	0.74	0.80	0.76
'array'	NB	0.69	0.64	0.69	0.65
	RB	0.76	0.71	0.76	0.73
'low'	NB	0.20	0.15	0.20	0.14
	RB	0.14	0.05	0.14	0.07
'very low'	NB	0.05	0.03	0.05	0.03
	RB	0.04	0.01	0.04	0.02

Acc: accuracy, Pr: precision*, Rec: recall*, F1*
*weighted averages
NB: Nucleotide-based representation
RB: Reference-based representation
classification metrics are derived from 2d-labels

In Table 3 we present global accuracy and weighted averages for classification metrics precision, recall and F1 for the two data representations. The results for 2d-labels are shown and are available for the other labeling in Apendix section II. in both cases. We see that all data classes with the exception of genotyping arrays show higher classification metrics when the input to KNN is the integer-based DNA representation derived from the actual nucleotides at each position in the data matrix. At more than one step of the data pre-processing including the final merging of all data (See Methods section 3.4) we ensure that the all samples are encoded in the same manner by means of comparisons with the human reference. This tell us that this must not be a mistake where we for example flipped the reference nucleotide and then all encoded genotypes got a different integer representation. We also noticed 'low' and 'very low' data classes are in agreement with 'high' data, and the former two include within the cohorts both array-based

and NGS-based data at low coverage. From these observations we decided to use the nucleotide-based representation as starting input data for the following models and see the behaviour of 'array' data in light of this results.

The other most important finding from these results is the low efficiency of the model at classifying low-density data of the classes 'low' and 'very low', corresponding to aDNA data. Although this is not entirely surprising given the high levels of error and sparsity contained in it, we were hoping to achieve better metrics even in these preliminary models.

4.2 Impact of Standardization

Although every feature in our DNA sequences is a nucleotide among 4 possible and therefore all features are of the same nature, the integer-based representation that we used necessarily introduces quantitative differences among the nucleotides that do no actually exists and which KNN is not capable to separate . If we take the unique mapping used for the integer-based representation representation, A:1, C:2, G:3, T:4 and '.':5, it is implied by it that the missing state encoded by 5 is somehow closer to T than it is to A. This makes standardization necessary for the integer-based representation. We did it in the form of a simple scaling and centering transformation using mean subtraction and division by the standard deviation (See Methods section 3). We partitioned the data into training and testing samples following partition 'schema 1' and used the best hyperparameters ['algorithm': 'kd_tree', 'leaf_size': 10, 'n_neighbors': 1, 'p': 2, 'weights': 'distance'].

Table 4 shows also global accuracy and weighted averages for the classification metrics, this time for both standardized (centering and scaling; CS) and non-standardized data See Methods section 3.7.2. Again 2d-labels were employed for these comparisons and the complete results are available in Apendix section II.. We observe that accuracy goes from 0.89 to 0.93 for 'high' data. Interestingly, 'array' data shows a similar increase in accuracy but this is not higher than the 76% achieved based on RB representation with no standardization. Both 'low' and 'very low' data classes also experienced an increase on the classification metrics but remained overall below 20%, thus confirming the general lack of classification power of KNN given our data input. From these observations we decided to standardize the integer-based representation whenever suitable and importantly, we decided to analyze data classes 'high' and 'array' in separate from 'low' and 'very low' in an attempt to better capture the embedded structure within each of these data.

4.3 Analyzing aDNA data separately

Following previous observations we decided to analyze aDNA data separately, both 'low' and 'very low' data classes in order to assess if this had the effect of improving accuracy

Table 4. Impact of Standardization on KNN

Data class	Standardization	Acc.	Pr.	Rec.	F1
'high'	$-$	0.89	0.87	0.89	0.87
	CS	0.93	0.91	0.93	0.92
'array'	$-$	0.69	0.64	0.69	0.65
	CS	0.74	0.71	0.76	0.73
'low'	$-$	0.20	0.15	0.20	0.14
	CS	0.19	0.13	0.19	0.13
'very low'	$-$	0.05	0.03	0.05	0.03
	CS	0.03	0.01	0.03	0.02

Acc: accuracy, Pr: precision*, Rec: recall*, F1*
*weighted averages
CS: centering and scaling
classification metrics are derived from 2d-labels

in general but more specifically on the aDNA data. This decision followed the rational that apart from the high levels of genotyping error, aDNA contains genetic variation that never made it into the present and thus is sometimes simply not represented in data based on modern individuals ('high' and 'array' data classes). This might result in slightly - or even significantly - different embedding data representations between aDNA and modern data such as 'array' and 'high' data classes. To analyze the data in separate, we used a second partitioning strategy referred to as 'schema 2'. This partitioning follows the same logic as schema 1 but training and testing sets are made first for 'high' and 'array' data classes together, KNN model is fit, predictions are made and classification metrics are obtained. Then we repeated the same procedure but this time with data classes 'low' and 'very low' exclusively (See Methods section 3). The best hyperparameters used for this model were ['algorithm': 'kd_tree', 'leaf_size': 10, 'n_neighbors': 1, 'p': 2, 'weights': 'distance'].

We can see from Table 5 that both data classes 'high' and 'low' show an improvement in accuracy compared to classification using the partitioning schema 1 as shown in the previous section. 'very low' data shows no significant change. Interestingly, 'array' data showed a different trend once again with global accuracy dropping to the lowest for the models so far here described. We have included in Table 5 classification metrics based on the lowest resolution possible for a haplogroup which is the alphabetic letter of the main branch it belongs to. This shows that not even at this resolution level classification

Table 5. Modern vs aDNA data

Data class	Label	Acc.	Pr.	Rec.	F1
'high'	$2d$	0.94	0.92	0.94	0.93
	$main\ branch$	0.99	0.99	0.99	0.99
'array'	$2d$	0.72	0.65	0.72	0.68
	$main\ branch$	0.98	0.98	0.98	0.98
'low'	$2d$	0.22	0.13	0.22	0.15
	$main\ branch$	0.36	0.45	0.36	0.39
'very low'	$2d$	0.04	0.03	0.04	0.03
	$main\ branch$	0.10	0.30	0.10	0.15

Acc: accuracy, Pr: precision*, Rec: recall*, F1*
*weighted averages

shows good performance for either type of aDNA data ('low' and 'very low'). It also shows that predictions at this resolution have a very high degree of certainty when array data is used. This might have good practical use in indeed these results can be extended to other low-resolution array platforms for which other methods perform poorly given the set of positions they consider into the analysis.

it is often the case in aDNA studies that in order to produce PCA plots of the data in combination with modern individuals, rather than calculating eigenvectors based on the combined data set, the main axis of variation are first computed based on the modern data and aDNA data is then projected onto those main axis of variation. Invariably it is observed that aDNA data lives in different sub-spaces of variation from those of modern data. However these type of analysis are based on the complete genome where recombination exists and bigger differences accumulate over time. These differences are not the case for the the non-recombining Y chromosome, where even now extinct haplogroup lineages still exist as an embedded vector of variation within modern samples. Following this, we attempted to fit a KNN model based on 'high' and 'array' data classes and then use it to classify 'low' and 'very low' aDNA data. This approach however did not result in improved global accuracy for aDNA data Apendix section II..

From the combined observations here presented, we concluded that the biggest potential for this type of approach is likely the classification of Y chromosome haplogroups based on data from modern samples and genotyped using the various low-resolution arrays available. As a next step in this investigation, we propose a more focused effort on tackling classification using KNN and array genotyping data using standardize data from

the reference-base representations together with the panel of aDNA data (which contains itself array-based information but with higher levels of sparsity since not all positions in the chip get successfully typed) which has shown here to increase efficiency when used in the training of the KNN models. Besides this line of further investigation and technical development we also wanted to assess the feasibility of a novel application to DNA variation analysis based on words of nucleotides represented by numerical embeddings. We discuss briefly about it in the next sections and also summarize the observations encountered when using PCA as a dimensionality reduction technique as an intermediate step between our data transformations and its input to KNN.

4.4 Dimension reduction techniques (PCA)

The Y chromosome genetic data that we work with here is highly dimensional (65,177 nucleotide-based features). It is likely that most of the explained variation in the data can be summarized into a lower number of dimensions with little information loss. We wanted to asses and measure the effect in terms of accuracy change. We applied PCA on already standardize data and use all the principal components obtained (See Methods section 3.7.2) as downstream input for the KNN classifier. We applied this transformation in all cases presented before and in all of them a slight reduction in accuracy is observed compared to the same approach without PCA as an intermediate step (See Apendix section II).

4.5 Word-based embeddings

We wanted to further explore the possibility of using a more sophisticated dimensional reduction technique that could have a gain in accuracy while still reducing the complexity of the feature space. We decided to use a novel approach to the problem and use Wor2vec representation where 'words' were formed out of contiguous nucleotides in our Y chromosome-based data matrix and embeddings containing contextual information from neighboring genetic regions (adjacent words) were calculated in the form of numerical vectors of reduced dimension (See Methods section 3). Unfortunately this approach has not resulted yet in the expected improved accuracy we wanted. Given the efforts in developing and refining the conceptual framework together with the actual implementation, we will continue to work on this approach. Ideally the results of this can be extended to other problems in the field of genomics and bioinformatics, by providing an efficient and dense numerical representation of genomic sequence that incorporates contextual properties that can be exploited in other applications. We make available the implementation using TensorFlow in the Apendix section II.

5 Conclusion

- The poor performance that we observe in both types of aDNA data (low and very low coverage) makes us conclude that KNN is not the ideal approach to Y chromosome haplogroup classification based on this type of data, even though we make good efforts to only include samples annotated as having good quality measures and the labels included have some level of manual curation in many cases. Other ML approaches and data prepossessing procedures might be further investigated.

- KNN performs quite well given high-density data. A reduced number of poorly samples haplogroup classes drive the global accuracy somewhat lower than we would like to see, however for many haplogroup classes prediction is very accurate. Nevertheless, there are several methods for classification based on high-density data so this is not the exact candidate as data type to classify based on this approach.

- We have gather good evidence in order to develop further this approach for the classification of array-based data. This type of sparse data is often difficult for available prediction methods (not designed for that) and our approach does good at classifying it. The trained model is available to the public and we plan to work further on this direction

- Finally we also plan to work further on the potential applications of word-based numerical embeddings based on genetic sequences. This includes its use as input for this and other types of machine learning algorithms. Not only for the case presented here of haplogroup prediction but also as a generator of genetic sequence with similar contextual properties as those observed in the genome.

Acknowledgements

I would like to thank all the people that supported me to pursue this masters degree and who contributed so that I could walk through it smoother. I specially thank Dr. Kallol Roy and Dra. Monika Karmin for their willingness and all good spirit to work on this project together.

References

Athey, T Whit (2006). "Haplogroup Prediction from Y-STR Values Using Using an Allele- Frequency Approach". In: *Journal of Genetic Genealogy* 2.2, pp. 34–39. ISSN: 15573796.

Barupal, Dinesh Kumar and Oliver Fiehn (2019). "Generating the blood exposome database using a comprehensive text mining and database fusion approach". In: *Environmental Health Perspectives* 127.9, pp. 2825–2830. ISSN: 15529924. DOI: 10.1289/EHP4713.

Behar, Doron M. et al. (2008). "The Dawn of Human Matrilineal Diversity". In: *American Journal of Human Genetics* 82.5, pp. 1130–1140. ISSN: 00029297. DOI: 10.1016/j.ajhg.2008.04.002. URL: http://dx.doi.org/10.1016/j.ajhg.2008.04.002.

Bergström, Anders et al. (2020). "Insights into human genetic variation and population history from 929 diverse genomes". In: *Science* 367.6484. ISSN: 10959203. DOI: 10.1126/science.aay5012.

Brennan, Paul M. et al. (2017). "Pre-operative obesity does not predict poorer symptom control and quality of life after lumbar disc surgery". In: *British Journal of Neurosurgery* 31.6, pp. 682–687. ISSN: 1360046X. DOI: 10.1080/02688697.2017.1354122.

Byrska-Bishop, Marta et al. (2021). "High Coverage Whole Genome Sequencing of the Expanded 1000 Genomes Project Cohort Including 602 Trios". In: *bioRxiv*. DOI: 10.1101/2021.02.06.430068.

Capelli, Cristian et al. (2006). "Population structure in the Mediterranean basin: A Y chromosome perspective". In: *Annals of Human Genetics* 70.2, pp. 207–225. ISSN: 14691809. DOI: 10.1111/j.1529-8817.2005.00224.x.

Casanova, Myrum et al. (1985). "A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance". In: *Science* 230.4732, pp. 1403–1406. ISSN: 00368075. DOI: 10.1126/science.2999986.

Casella, G, S Fienberg, and I Olkin (2017). *An Introduction to Statistical Learning*. Springer. ISBN: 9781461471370.

Hallast, Pille et al. (2021). "A common 1.6 mb Y-chromosomal inversion predisposes to subsequent deletions and severe spermatogenic failure in humans". In: *eLife* 10, pp. 1–22. ISSN: 2050084X. DOI: 10.7554/eLife.65420.

Hammer, Michael (2002). "A nomenclature system for the tree of human Y-Chromosomal binary haplogroups". In: *Genome Research* 12.2, pp. 339–348. ISSN: 10889051. DOI: 10.1101/gr.217602.

Jobling, M. A., A. Pandya, and C. Tyler-Smith (1997). "The Y chromosome in forensic analysis and paternity testing". In: *International Journal of Legal Medicine* 110.3, pp. 118–124. ISSN: 09379827. DOI: 10.1007/s004140050050.

Jobling, Mark et al. (2014). *Human Evolutionary Genetics: 2nd edition*. Vol. 76. 6, pp. 134–164. ISBN: 978-0-8153-4148-2.

Jostins, Luke et al. (2014). "YFitter: Maximum likelihood assignment of Y chromosome haplogroups from low-coverage sequence data". In: *BioRxiv*, pp. 1–6. URL: http://arxiv.org/abs/1407.7988.

Karmin, Monika et al. (2015). "A recent bottleneck of Y chromosome diversity coincides with a global change in culture". In: *Genome Research*, pp. 1–8. ISSN: 15495469. DOI: 10.1101/gr.186684.114.67. URL: http://genome.cshlp.org/content/25/4/459.full.pdf.

Lazaridis, Iosif et al. (2014). "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518, pp. 409–413. ISSN: 0028-0836. DOI: 10.1038/nature13673. URL: http://www.nature.com/doifinder/10.1038/nature13673.

Mallick, Swapan et al. (2016). "The Simons Genome Diversity Project: 300 genomes from 142 diverse populations". In: *Nature* 538.7624, pp. 201–206. ISSN: 0028-0836.

Martiniano, R. et al. (2020a). "Placing ancient DNA sequences into reference phylogenies". In: *bioRxiv*, pp. 1–19. ISSN: 26928205. DOI: 10.1101/2020.12.19.423614.

— (2020b). "Placing ancient DNA sequences into reference phylogenies". In: *bioRxiv*, pp. 1–19. ISSN: 26928205. DOI: 10.1101/2020.12.19.423614.

Metzker, Michael L. (2010). "Sequencing technologies the next generation". In: *Nature Reviews Genetics* 11.1, pp. 31–46. ISSN: 14710056. DOI: 10.1038/nrg2626. URL: http://dx.doi.org/10.1038/nrg2626.

Mikolov, Tomas et al. (2013). "Efficient estimation of word representations in vector space". In: *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pp. 1–12.

Poznik, G David (2016a). "yHaplo: Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men". In: pp. 1–6.

Poznik, G David et al. (2016). "Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences". In: *Nature Genetics* 12.9, pp. 809–809. ISSN: 1061-4036. DOI: 10.1038/ng.3559. URL: http://www.nature.com/doifinder/10.1038/ng.3559.

Poznik, G. David (2016b). "Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men". In: *bioRxiv*, p. 088716. DOI: 10.1101/088716. URL: https://www.biorxiv.org/content/early/2016/11/19/088716.

Ralf, Arwin et al. (2018). "Yleaf: software for human Y-chromosomal haplogroup inference from next generation sequencing data". In: *Molecular Biology and Evolution* March, pp. 1–4. ISSN: 0737-4038. DOI: 10.1093/molbev/msy032. URL: https://academic.oup.com/mbe/advance-article/doi/10.1093/molbev/msy032/4922696.

Reich, David, Alkes L. Price, and Nick Patterson (2008). "Principal component analysis of genetic data". In: *Nature Genetics* 40.5, pp. 491–492. ISSN: 10614036. DOI: 10.1038/ng0508-491.

Schlecht, Joseph et al. (2008). "Machine-learning approaches for classifying haplogroup from Y chromosome STR data". In: *PLoS Computational Biology* 4.6. ISSN: 1553734X. DOI: 10.1371/journal.pcbi.1000093.

Severson, Alissa L. et al. (2018). "SNAPPY: Single Nucleotide Assignment of Phylogenetic Parameters on the Y chromosome". In: *bioRxiv*. DOI: 10.1101/454736. URL: https://www.biorxiv.org/content/early/2018/10/29/454736.1.

Underhill, Peter A. and Toomas Kivisild (2007). "Use of Y Chromosome and Mitochondrial DNA Population Structure in Tracing Human Migrations". In: *Annual Review of Genetics* 41.1, pp. 539–564. ISSN: 0066-4197. DOI: 10.1146/annurev.genet.41.110306.130407. URL: http://www.annualreviews.org/doi/10.1146/annurev.genet.41.110306.130407.

Wong, Emily H.M. et al. (2017). "Reconstructing genetic history of Siberian and Northeastern European populations". In: *Genome Research* 27.1, pp. 1–14. ISSN: 15495469. DOI: 10.1101/gr.202945.115.

Wrighton, Katharine (2021). "Filling in the gaps telomere to telomere". In: *Nature Research 2021* 1162.February, p. 2021. URL: https://www.nature.com/articles/d42859-020-00117-1.

Appendix

I. Glossary

allele. An allele is one of two or more versions of DNA sequence (a single base or a segment of bases) at a given genomic location.

assignment. The allocation of someone or something as belonging to a particular group or category. The terms assignment, classification and labeling are often used interchangeably in this text

clade. A group of organisms believed to comprise all the evolutionary descendants of a common ancestor.

classification. Classification is the process of categorizing a given set of data into classes. See the definition of assignment also.

coverage. Next-generation sequencing (NGS) coverage describes the average number of reads that align to, or "cover," known reference bases.

genotype. A genotype is a scoring of the type of variant present at a given location (i.e., a locus) in the genome. It can be represented by symbols. For example, A, C, G, T.

haplogroup. A haplotype is a group of alleles in an organism that are inherited together from a single parent and a haplogroup is a group of similar haplotypes that share a common ancestor. The non-recombining portion of the Y chromosome is essentially a long haplotype.

high-density data. In this work, the term high-density data refers to genetic data from an individual for which a high number of positions have been assayed and the specific genotype at each of these loci (plural of locus) is known.

INDEL. Is a molecular biology term for an insertion or deletion of bases in the genome of an organism.

label. A label in the context of machine learning is a category that we are trying to predict. See the definition of assignment.

low-density data. In this work, the term low-density data refers to genetic data from an individual for which only a reduced and sometimes arbitrary subset of variation has

been assayed or is only available due to methodological reasons. This term is used interchangeably with sparse genetic data.

monophyletic taxon. A group composed of a collection of organisms, including the most recent common ancestor of all those organisms and all the descendants of that most recent common ancestor. A monophyletic taxon is also called a clade.

NGS. Next-generation sequencing (NGS) is a massively parallel sequencing technology that offers ultra-high throughput, scalability, and speed. The technology is used to determine the order of nucleotides in entire genomes or targeted regions of DNA or RNA.

phylogenetic tree. A phylogenetic tree, also known as a phylogeny, is a diagram that depicts the lines of evolutionary descent of different species, organisms, or genes from a common ancestor.

phylogeography. Phylogeography is the study of the historical processes that may be responsible for the past to present geographic distributions of genealogical lineages. This is accomplished by considering the geographic distribution of individuals in light of genetics, particularly population genetics.

prediction. In this work the term prediction is used interchangeably with the terms classification, assignment and labeling.

SNP. A single-nucleotide polymorphism is a germline substitution of a single nucleotide at a specific position in the genome. The historical definition of SNPs requires the substitution to be present in a sufficiently large fraction of the population (e.g. 1% or more)

SNV. A DNA sequence variation that occurs when a single nucleotide (adenine, thymine, cytosine, or guanine) in the genome sequence is altered. Although not exactly the same, in this work the term is used interchangeably with SNP.

sparse data. See the definition of low-density data.

targeted sequencing. Targeted sequencing is a rapid and cost-effective way to detect known and novel variants in selected sets of genes or genomic regions.

variant. An alteration in the most common DNA nucleotide sequence (e.g. SNPs).

WGS. Whole genome sequencing (WGS), also known as full genome sequencing, com-

plete genome sequencing, or entire genome sequencing, is the process of determining the entirety, or nearly the entirety, of the DNA sequence of an organism's genome at a single time.

Y-STR markers. Short tandem repeats (STRs) are short repeated sequences of DNA (2–6 bp) that account for approximately 3% of the human genome. Y-STR markers occur in the Y chromosome.

Y-karyotype. SA karyotype is a preparation of the complete set of metaphase chromosomes in the cells of a species or in an individual organism, sorted by length and other features.

recombination. Genetic recombination is the exchange of genetic material between different organisms which leads to production of offspring with combinations of traits that differ from those found in either parent

II. Code and supplementary materials

Code and all supplementary material related to this work can be found at https://github.com/JRodrigoF/yhapML

III. Licence

Non-exclusive licence to reproduce thesis and make thesis public

I, **Jose Rodrigo Flores Espinosa**,
*(*author's name)

1. herewith grant the University of Tartu a free permit (non-exclusive licence) to

   reproduce, for the purpose of preservation, including for adding to the DSpace digital archives until the expiry of the term of copyright,

   "Classification of human Y chromosome haplogroups based on dense and sparse genetic data using machine learning approaches",
   *(*title of thesis)

   supervised by Dr. Kallol Roy and Dra. Monika Karmin.
   *(*supervisor's name)

2. I grant the University of Tartu a permit to make the work specified in p. 1 available to the public via the web environment of the University of Tartu, including via the DSpace digital archives, under the Creative Commons licence CC BY NC ND 3.0, which allows, by giving appropriate credit to the author, to reproduce, distribute the work and communicate it to the public, and prohibits the creation of derivative works and any commercial use of the work until the expiry of the term of copyright.

3. I am aware of the fact that the author retains the rights specified in p. 1 and 2.

4. I certify that granting the non-exclusive licence does not infringe other persons' intellectual property rights or rights arising from the personal data protection legislation.

J. Rodrigo Flores E.
17/05/2022