Forensic Science International: Genetics xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs

A. Van Geystelen [a,b,1], R. Decorte [a,c], M.H.D. Larmuseau [a,c,d,1,*]

[a] UZ Leuven, Laboratory of Forensic Genetics and Molecular Archaeology, Leuven, Belgium
[b] KU Leuven, Department of Biology, Laboratory of Socioecology and Social Evolution, Leuven, Belgium
[c] KU Leuven, Department of Imaging & Pathology, Forensic Medicine, Leuven, Belgium
[d] KU Leuven, Department of Biology, Laboratory of Biodiversity and Evolutionary Genomics, Leuven, Belgium

ARTICLE INFO

Article history:

Keywords:
Haploid markers
Y-chromosome
Phylogenetic tree
Bio-informatics
Whole-genome SNP calling
Y-SNPs

ABSTRACT

The Y-chromosomal phylogenetic tree has a wide variety of important forensic applications and therefore it needs to be state-of-the-art. Nevertheless, since the last 'official' published tree many publications reported additional Y-chromosomal lineages and other phylogenetic topologies. Therefore, it is difficult for forensic scientists to interpret those reports and use an up-to-date tree and corresponding nomenclature in their daily work. Whole genome sequencing (WGS) data is useful to verify and optimise the current phylogenetic tree for haploid markers. The AMY-tree software is the first open access program which analyses WGS data for Y-chromosomal phylogenetic applications. Here, all published information is collected in a phylogenetic tree and the correctness of this tree is checked based on the first large analysis of 747 WGS samples with AMY-tree. The obtained result is one phylogenetic tree with all peer-reviewed reported Y-SNPs without the observed recurrent and ambiguous mutations. Nevertheless, the results showed that currently only the genomes of a limited set of Y-chromosomal (sub-)haplogroups is available and that many newly reported Y-SNPs based on WGS projects are false positives, even with high sequencing coverage methods. This study demonstrates the usefulness of AMY-tree in the process of checking the quality of the present Y-chromosomal tree and it accentuates the difficulties to enlarge this tree based on only WGS methods.

1. Introduction

A state-of-the-art phylogeny of the human Y-chromosome based on bi-allelic polymorphisms is an essential tool for forensic genetics. Forensic scientists are taking advantage of the Y-chromosomal phylogenetic tree in their daily work, e.g. by checking the quality of datasets or by assigning geographical landscapes to specific lineages [1,2]. Y-chromosomal single nucleotide polymorphisms (Y-SNPs) have a great capacity for detecting geographical origins as many lineages defined by Y-SNPs show a strong continent-specific [3,4] and even intra-continent-specific distribution [5–7]. Their usefulness is illustrated by the fact that Y-SNP data are now also included in Y-chromosomal forensic databases, such as in the YHRD database [8]. Therefore, an up-to-date extended Y-chromosomal phylogeny based on these bi-allelic markers which are preferably unambiguous and non-recurrent but which have a high discrimination power is required for forensic applications.

Since the publication of the latest 'official' Y-chromosomal phylogenetic tree by Karafet et al. [4], a continuous wave of new peer-reviewed articles which report changes to this tree are published. These changes include a new root and new basal clades [9,10], modifications of the global backbone [3,11], different phylogenetic topologies within a haplogroup [12–14], newly described sub-haplogroups [15–17], or other phylogenetic positions for a certain mutation [18]. As these publications are not coordinated different names are given to the Y-chromosomal lineages for which the phylogenetic position is given in different topologies. Therefore, the currently overall reported Y-chromosomal tree is not clear and this makes it difficult for forensic researchers to use a uniform phylogenetic tree. Hence new initiatives to ensure more continuity in the report of the most recent phylogenetic Y-chromosomal tree are needed.

Large whole genome sequencing (WGS) projects such as the 1000 Genomes Project [19,20] bring an opportunity to introduce the required uniformity in the reporting of the haploid

* Corresponding author at: Katholieke Universiteit Leuven, Forensic Medicine, Kapucijnenvoer 33, B–3000 Leuven, Belgium. Fax: +32 0 16324575.
E-mail address: maarten.larmuseau@bio.kuleuven.be (M.H.D. Larmuseau).
[1] Both authors contributed equally to this study.

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.03.010

Y-chromosomal tree. The analysis of whole Y-chromosomes within male genomes allows verification and optimisation of the currently used phylogenetic tree. WGS data has already proved to be useful in verifying and optimising the phylogeny of the other haploid markers in the human genome i.e. the mitochondrial DNA (mtDNA). Relevant ambiguous markers and back-mutations which influence the interpretation of previous forensic and evolutionary genetic studies were detected based on these data [21]. Recently a new Y-chromosomal phylogenetic tree was built after a *tabula rasa* of the present Y-chromosomal tree by using only Y-SNPs from available WGS male samples [22]. By comparing this new phylogenetic tree with the currently used one, the backbone of the currently used phylogenetic tree was confirmed in this study. However, this new tree is not useful for forensic research because there is no link between currently used and newly reported lineages. Furthermore, the set of used genomes is not a good representation of all existing Y-chromosomal (sub-)haplogroups and geographical regions. There are also still too much false positive SNP calls in this WGS dataset. Alternatively, the AMY-tree software is the first open access program which academics and forensic professionals can use to verify and optimise the currently used Y-chromosomal tree by using WGS data [23]. The first AMY-tree analysis was done based on 118 WGS samples and proved already its usefulness to verify and to optimise the present Y-chromosomal phylogenetic tree [23].

The aim of this study is to perform the largest reported screening of male genomes for Y-chromosomal phylogenetic applications based on the AMY-tree software. Firstly, we want to merge all newly Y-SNPs from recent peer-reviewed publications since the latest 'official' Y-chromosomal phylogeny [4] into one single tree which is useful for forensic applications. Secondly, this updated Y-chromosomal phylogenetic tree needs to be checked for recurrent mutations, ambiguous SNPs and other difficulties for the (forensic) application of the tree. Thirdly, this study also wants to find out for which Y-chromosomal (sub-)haplogroups there is already WGS data available. Finally, investigating the possibilities to enlarge the Y-chromosomal phylogenetic tree based on the current Y-SNP detections in WGS data is the last aim of this study.

2. Materials and methods

2.1. Updated phylogenetic tree

The latest updated phylogenetic tree of the Y-chromosome as it was published by Van Geystelen et al. [23] was manually updated based on recent descriptions of new Y-SNPs in academic research papers like Pamjav et al. [16] and Scozzari et al. [9]. As the exact phylogenetic position of a few new Y-SNPs was not given their position needed to be determined based on the results of AMY-tree of all WGS samples. Next, also recurrent mutations, ambiguous SNP-loci and wrongly defined mutation conversions within the newly updated Y-chromosomal tree were ascertained based on those AMY-tree results.

2.2. WGS Y-SNPs dataset

In order to check the manually updated phylogenetic tree and to optimise the AMY-tree software, a large dataset of whole genome Y-SNP calls was assembled. This dataset consists of 747 samples which represent 660 males, as several genomes were analysed in different projects. Within this dataset the genomes of eight males whose father's genome was also sequenced are present. The SNP calls were collected from four large WGS projects and several individual genome projects (Supplementary Materials Table S1). These projects differ from each other based on the used next-generation sequencing (NGS) platforms and sequence coverage. First, Complete Genomics made the SNP calls of 35 whole genomes of males available (http://www.completegenomics.com/public-data/69-Genomes/04 Jan 2013); those genomes were sequenced with a high sequencing coverage on the Complete Genomics Analysis (CGA) Platform [24]. Second, the Personal Genome Project (PGP) and Singapore Sequencing Malay Project (SSMP) also used this CGA platform. PGP is a project started to obtain and openly share human genome sequences in combination with health information. At the moment 40 male genomes were available (www.personalgenomes.org 04 Jan 2013). The SSMP on the other hand wanted to characterise the polymorphic variants in the population of Malays, an Austronesian group present in Southeast Asia and Oceania. Recently, the Y-SNP calls of 46 Malays were made publically available [25]. Next, the 1000 Genomes Project aims to provide a comprehensive resource on human genetic variation by sequencing more than 1000 human genomes. In 2010, SNP calls of 77 males were made available in the pilot phase [20] and two years later a set of 526 SNPs profiles were published as result of phase 1 of the Project [19]. As the 1000 Genomes project aims to sequence a large number of people, the sequencing coverage was lower than in the other projects. Finally, 23 additional samples were collected from several single genome projects [26–35, 36 and unpublished genomes of Guy Froyen].

2.3. AMY-tree modifications

Several modifications to the AMY-tree software version 1.0 [23] were made for the assessment of the SNP calling quality in WGS data. This was necessary as the quality of the SNP calling influences the AMY-tree analysis of a sample and therefore also the interpretation of the result of the analysis [23]. The extra quality assessment is based on the results of the first AMY-tree run of a certain sample. This assessment assumes that the used phylogenetic tree is correct and that the assigned haplogroup after the first AMY-tree run is the actual haplogroup of that sample.

The algorithm for the extra quality assessment is simple and comprehensible as shown in Fig. 1. First, all Y-SNPs of the phylogenetic tree are selected except if the determined haplogroup of the first run is a paragroup (e.g. R1b1b2*). In the case of a paragroup, all Y-SNPs which are in sub-nodes of the main group of this paragroup are excluded from the selection. This is done in order to remove the influence of too much false positive SNP calls

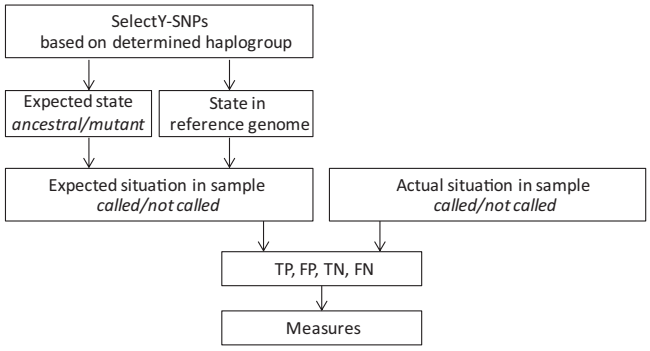

Fig. 1. Workflow of the quality assessment algorithm of the AMY-tree, version 1.1. First, certain Y-SNPs of the Y-chromosomal phylogenetic tree are selected based on the determined haplogroup of the first run in order to avoid too much false positive Y-SNPs. Next, the expected state (*ancestral* or *mutant*) of the selected SNPs is determined based on the haplogroup and the phylogenetic tree. These expectations of state are converted to expectations of *called* or *not called* based on the SNP state in the reference genome. These expectations are then compared to the actually called SNPs of the sample such that the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) will be determined. Finally, several different measures will be calculated.

Please cite this article in press as: A. Van Geystelen, et al., Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs, Forensic Sci. Int. Genet. (2013), http://dx.doi.org/10.1016/j.fsigen.2013.03.010