SNAPPY: Single Nucleotide Assignment of Phylogenetic Parameters on the Y chromosome

Alissa L. Severson[1,*], Jonathan A. Shortt[2,*], Fernando L. Mendez[1], Genevieve L. Wojcik[1], Carlos D. Bustamante[1], Christopher R. Gignoux[2,3]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA;

[2]Colorado Center for Personalized Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO 80045, USA

[*]Contributed equally

[3]To whom correspondence should be addressed

Abstract

Summary

The assignment of Y chromosome data to related clusters, or haplogroups, is a common application in human population genetics. To enable this at scale, we developed SNAPPY. SNAPPY is a software program used to assign Y-chromosome phylogeny-informed haplotypes using dense genotype data. The program efficiently tests all haplotypes in a provided Y-chromosome database to find the haplogroup that is best supported by the input genotypes. Importantly, the method considers both the amount of support for the specific haplogroup, as well as its ancestral haplogroups via parsimony. This accounts for the underlying genealogy the haplotypes represent, strengthening the accuracy of the assignments. SNAPPY is fast, scalable, and uses standard file formats, making it easy to integrate into analytical pipelines.

Availability and Implementation

The program is implemented in python. The program, a user manual, haplotype databases, and test datasets are available for download at github.com/chrisgene/snappy.

Contact

Jonathan.shortt@ucdenver.edu, Chris.gignoux@ucdenver.edu

Introduction

Analyses of Y-chromosome haplogroups have yielded important insights into human migration history (Bergstrom, et al., 2016; Chiaroni, et al., 2009), cultural customs (Seielstad, et al., 1998), and ancestral population sizes (Karmin, et al., 2015). These insights are possible because of the two unique characteristics of the Y-chromosome: it is passed directly from father to son, and much of the Y-chromosome does not recombine with any other chromosome; this leaves a long tract of unbroken DNA that serves as faithful record of evolution of the chromosome throughout human history. Constant accumulation of variation within the chromosome over human history created new combinations of variants (haplotypes), which trace patrilineal descent.

Despite its use in ancestry, efficient software tools that are able to rapidly assign Y-chromosome haplotypes from dense genotype data into clusters of related haplotypes (haplogroups) at population scales are limited in scope. Here we describe SNAPPY (Single Nucleotide Assignment of Phylogenetic Parameters on the Y-chromosome), a tool to assign Y-chromosome haplogroups using dense genotyping data. Haplogroup assignment is based on the well-established polymorphisms along the Y-chromosome with haplogroup information

maintained by the International Society of Genetic Genealogists (ISOGG, isogg.org) and others [e.g., (Karafet, et al., 2015; Poznik, et al., 2016)], using only phylogenetically-informative alleles to determine which haplogroup has the highest support, thus avoiding complications related to the reversion of alleles to ancestral states. Importantly, the method is able to identify haplogroups from both leaf and interior nodes of the Y-chromosome tree. Here we briefly outline the implementation of the program, and present Y-chromosome haplotype assignments from The 1000 Genomes Project (Genomes Project, et al., 2015) data to validate the algorithm.

Implementation

Dependencies

SNAPPY is implemented in python requiring only numpy in addition to the standard python library. In addition, SNAPPY makes use of plink (Chang, et al., 2015) for initial file format conversions from either array-based formats (e.g., .ped/.bed) or from vcf.

Algorithm Description

SNAPPY leverages the Y-chromosome phylogeny and a database of haplotype-informative SNPs derived from ISOGG and other sources to store nested haplotypes in memory-efficient dictionaries. Genotypes from individual samples are stored as a list of dictionaries keyed by chromosome position. Currently the tree is optimized for sites on the commonly-used Illumina Multi-Ethnic Global Array (pagestudy.org/mega), however tree files can easily be generated from relevant sources. To compute haplogroup scores for each of the 565 possible Y-chromosome haplogroups present in our Y-chromosome haplogroup reference library, SNAPPY looks up a sample's genotypes—stored in a dictionary—and counts the proportion of matching alleles that are at haplogroup-informative positions.

63 SNAPPY determines a score for each haplotype using the number of haplogroup-informative
64 derived alleles adjusted by the sample's number of non-missing informative positions. Note that
65 a particular haplogroup's score uses alleles from both its own informative positions as well as its
66 ancestral nodes on the tree. This ensures that haplogroup assignments take into account the full
67 phylogenetic structure of the Y-chromosome, and enables the creation of easily traversable data
68 structures that eliminate redundant storage of informative positions due to the highly
69 parsimonious nature of variants on the Y(Poznik, et al., 2016). Haplogroup scores for every
70 haplogroup are stored in a two-dimensional numpy array to allow for efficient storage and quick
71 processing.

72 Finally, haplogroup assignments are made by evaluating the scores of nodes of each
73 branch, starting at the most distal node of the branch that has both a score above a user-defined
74 threshold and no descendant haplotypes with a score that exceeds the user-defined threshold, and
75 then working towards the root of the tree. This ensures that all nodes are considered potential
76 haplogroups for the sample, even if they are not terminal nodes on the tree. SNAPPY makes its
77 haplogroup assignment based on the highest scoring node or the deepest node with a score higher
78 than a user-defined threshold. This algorithm ensures that the deepest haplogroup with sufficient
79 support is assigned. In addition to reporting the single haplogroup with the most support for each
80 sample, SNAPPY also reports each haplogroup that has a score greater than a user-defined
81 minimum score so that haplogroup assignments can be adjusted or investigated where necessary.

Validation and Testing

Data Sources

We downloaded a list of Y-chromosome variants found on the Multi-Ethnic Genotyping Array (MEGA) (Bien, et al., 2016) (a list of variants found on the MEGA can be found at https://pagestudy.org/index.php/multi-ethnic-genotyping-array). Because the MEGA Y content was designed to be ancestry informative, the SNPs included are well-distributed throughout the Y-chromosome tree, which allows for accurate and precise assignment of haplogroups. MEGA variant positions were converted from GRCh38.p7 to GRCh37 using NCBI's genome remapper tool (https://www.ncbi.nlm.nih.gov/genome/tools/remap) in preparation for extracting MEGA positions from Y-chromosome Phase 3 of The 1000 Genomes Project (Genomes Project, et al., 2015) data.

For reference, we downloaded 1,233 Y-chromosome genotypes from Phase 3 of The 1000 Genomes Project from NCBI (Genomes Project, et al., 2015) (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp). The vcf was filtered to contain only the variants that are present on MEGA using plink (Chang, et al., 2015). This yielded a final total of 2,366 variants.

Similarly, we provide Y-chromosome haplogroup tree information and informative SNPs, available in the SNAPPY distribution found at github.com/chrisgene/snappy. This information is largely consistent with phylogenetic information present on ISOGG.

Results

To assess the speed and accuracy of SNAPPY, we tested 1,233 males from Phase 3 of The 1000 Genomes Project (Genomes Project, et al., 2015). The program is computationally efficient: for a set of 1,233 samples genotyped at 2,366 SNPs, it runs in an average of 4.96 seconds on a single

2.3 GHz core of an Intel Xeon processor, using 278 Mb of RAM. SNAPPY scales linearly indicating good performance for larger data sets.

Y-chromosome haplotypes for 1000 Genomes males have been previously well-characterized (Poznik, et al., 2016), and we used this characterization to assess the accuracy of SNAPPY on genotype data compared to full sequences. We found that most individuals had top haplogroup scores >95%, (Supplemental Figure 1), correctly predicting over >99% of major haplogroup assignments for all individuals, with minor differences in fine-grained haplogroup designations, even given topological differences between our tree and prior examples (Supplemental Table 1). The three individuals with discordant major haplogroup assignments were assigned by SNAPPY to the haplogroup P1 rather than the closely-related reference assignment of Q1a. We note that the P1 and Q haplogroups had the same score (0.958) in these individuals, but the P1 haplogroup was chosen because it resides deeper in the tree than Q. These inconsistencies are expected to resolve with increased genotype density or tree topologies with improved resolution and can be easily adjudicated on manual inspection. A fourth individual was correctly assigned to the A0 haplogroup consistent with no derived haplogroup with support above our recommended 60% minimum match. Some differences between the sets are to be expected because the number of variants between the two datasets differ substantially (>60K for the reference set vs. 2,399 for our tests), and different Y-chromosome phylogenies (*de novo* reconstruction vs. ISOGG download) may result in subtle yet important differences between branch points. Nevertheless, the ability to recover major haplogroups from individuals representing nearly all major branches of the Y haplogroup tree indicates that SNAPPY, and informative array-genotyped sites, are robust and unbiased in ancestry assignment.

Conclusions

The ability to determine Y-chromosome haplogroups is of broad interest in many lines of research from population genetics to anthropology to medicine. Here we have introduced and demonstrated SNAPPY, a method to rapidly and accurately assign Y-chromosome haplogroups to population-scale data sets using dense genotyping. SNAPPY is user-friendly, requiring input in common plink format and just a single command to run. Additionally, SNAPPY is flexible, with several user-defined parameters, the ability to use user-curated y-haplogroup trees, and to manually interpret haplogroup assignments.

Funding Information

This work was supported in part by National Institutes of Health grant number T32HG00044 to CRG. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Acknowledgments

We gratefully acknowledge the community in ISOGG for maintaining and providing Y-chromosome haplogroup trees, David Poznik for assistance with the original phylogeny, and Peter Underhill for general discussions.

References

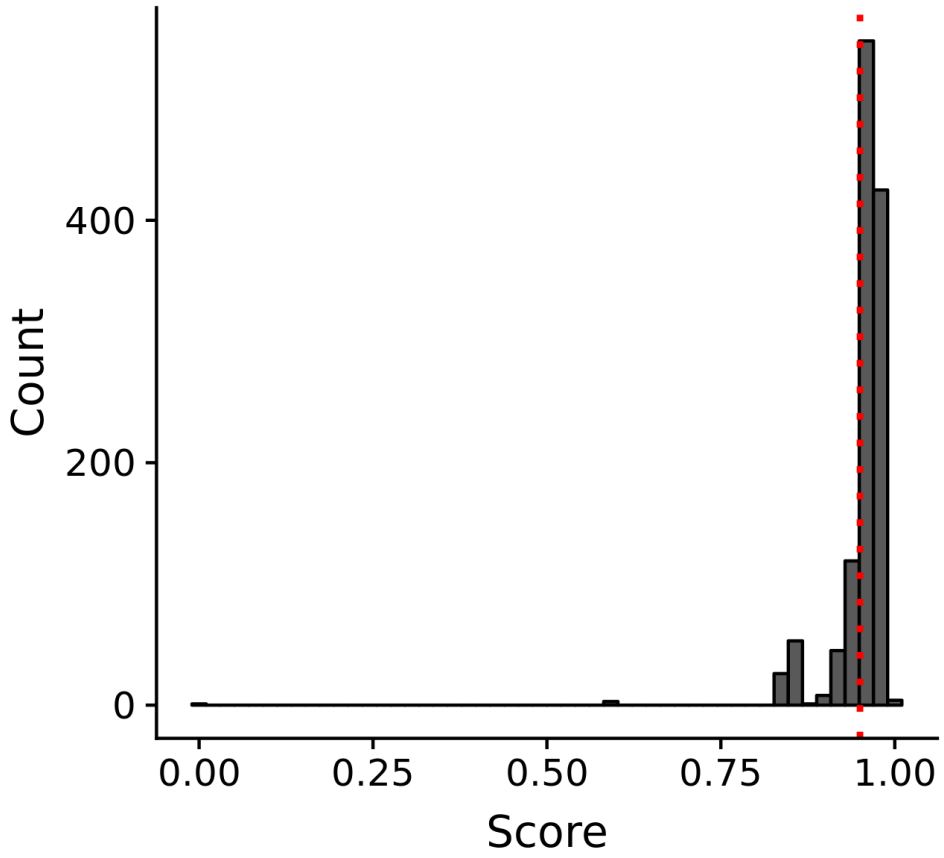
Bergstrom, A., *et al.* Deep Roots for Aboriginal Australian Y Chromosomes. *Curr Biol* 2016;26(6):809-813.

Bien, S.A., *et al.* Strategies for Enriching Variant Coverage in Candidate Disease Loci on a Multiethnic Genotyping Array. *PLoS One* 2016;11(12):e0167758.

Chang, C.C., *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 2015;4:7.

149 Chiaroni, J., Underhill, P.A. and Cavalli-Sforza, L.L. Y chromosome diversity, human
150 expansion, drift, and cultural evolution. *Proc Natl Acad Sci U S A* 2009;106(48):20174-20179.
151 Genomes Project, C.*, et al.* A global reference for human genetic variation. *Nature*
152 2015;526(7571):68-74.
153 Karafet, T.M.*, et al.* Improved phylogenetic resolution and rapid diversification of Y-
154 chromosome haplogroup K-M526 in Southeast Asia. *Eur J Hum Genet* 2015;23(3):369-373.
155 Karmin, M.*, et al.* A recent bottleneck of Y chromosome diversity coincides with a global
156 change in culture. *Genome Res* 2015;25(4):459-466.
157 Poznik, G.D.*, et al.* Punctuated bursts in human male demography inferred from 1,244
158 worldwide Y-chromosome sequences. *Nat Genet* 2016;48(6):593-599.
159 Seielstad, M.T., Minch, E. and Cavalli-Sforza, L.L. Genetic evidence for a higher female
160 migration rate in humans. *Nat Genet* 1998;20(3):278-280.

161



Supplemental Figure 1 Distribution of SNAPPY-assigned haplogroup scores. Histogram of haplogroup scores assigned by SNAPPY. A dashed red line indicates 95%.

Supplemental Table 1 Proportion of Concordant and Discordant SNAPPY Haplotype Assignments Relative to Major Reference Haplotypes

Reference Haplogroup Assignment	n	SNAPPY Haplogroup Assignment: A0	A1a	B	B2	C1	C2	C3	C5	D1	D2	E1a	E1b	E2	F	G1	G2	H	H0	H1	H2	HI	I1	I2	J1	J2	K	K2	L1	N	N1	N2	O1	O2	O3	P1	Q1a	Q1b	R1a	R1b	R2	T
A0	2	1																																							
A1a	3		1																																						
B	3			1																																					
B2	4				1																																			
C1	4					1																																		
C2	0																																							
C3	9						1																																	
C5	17					1																																		
D1	0																																							
D2	20									1																														
E1a	20											1																												
E1b	297												1																											
E2	5													1																										
F	1														1																										
G1	2															1																									
G2	17																1																								
H	7																	1																							
H0	12																					1																			
H1	40																			1																					
H2	2																			1																					
HI	0																																								
I1	31																						1																		
I2	17																							1																	
J1	15																								1																
J2	61																									1															
K	0																																								
K2	1																										1														
L1	27																											1													
N	7																												1												
N1	25																													0.96	0.04										
N2	0																																								
O1	17																															1									
O2	75																																1								
O3	111																																0.03	0.97							
P1	0																																								
Q1a	42																																			1					
Q1b	3																																		1						
R1a	84																																					1			
R1b	215																																						1		
R2	29																																							1	
T	8																																								1