Yleaf: Software for Human Y-Chromosomal Haplogroup Inference from Next-Generation Sequencing Data

Arwin Ralf,[1] Diego Montiel González,[1] Kaiyin Zhong,[1] and Manfred Kayser*,[1]
[1]Department of Genetic Identification, Erasmus MC University Medical Centre Rotterdam, Rotterdam, The Netherlands
*Corresponding author: E-mail: m.kayser@erasmusmc.nl.
Associate editor: Bing Su

Abstract

Next-generation sequencing (NGS) technologies offer immense possibilities given the large genomic data they simultaneously deliver. The human Y-chromosome serves as good example how NGS benefits various applications in evolution, anthropology, genealogy, and forensics. Prior to NGS, the Y-chromosome phylogenetic tree consisted of a few hundred branches, based on NGS data, it now contains many thousands. The complexity of both, Y tree and NGS data provide challenges for haplogroup assignment. For effective analysis and interpretation of Y-chromosome NGS data, we present Yleaf, a publically available, automated, user-friendly software for high-resolution Y-chromosome haplogroup inference independently of library and sequencing methods.

Key words: **human Y-chromosome, haplogroups, phylogeny, next-generation sequencing, NRY.**

In the time of NGS (or massively parallel sequencing, MPS), the amount of genomic data produced and made publically available is rapidly expanding, providing valuable resources for many areas of research and applications. Due to its haploid nature and male-specific inheritance, the nonrecombining part of the human Y-chromosome (NRY) is highly suitable for phylogenetic studies and for addressing questions in evolution, anthropology, population history, genealogy, and forensics (Jobling and Tyler-Smith 2017). Over recent years, NGS data allowed the phylogenetic NRY tree to dramatically increase in size and complexity (Hallast et al. 2015; Poznik et al. 2016). The two most comprehensive tree versions ISOGG (http://www.isogg.org/tree; last accessed May 1, 2017) and Yfull (https://www.yfull.com/tree; last accessed May 1, 2017) currently contain thousands of branches. However, the complexity of both, Y tree and NGS data provide immense challenges for NRY haplogroup assignment, which reflects a key element in many NRY applications. Here, we introduce Yleaf, a Phyton-based, easy-to-use, publically available software tool for effective NRY single nucleotide polymorphism (SNP) calling and subsequent NRY haplogroup inference from NGS data. By comparative whole genome data analysis, we demonstrate high concordance of Yleaf in NRY-SNP calling compared with well-established tools such as SAMtools/BCFtools (Li et al. 2009), and GATK (McKenna et al. 2010) as well as improved performance of Yleaf in NRY haplogroup assignment relative to previously developed tools such as clean_tree (Ralf et al. 2015), AMY-tree (Van Geystelen et al. 2013), and yHaplo (Poznik 2016).

Yleaf allows analyzing NRY sequence data from many types of NGS libraries, that is, whole genomes, whole exomes, large genomic regions, and large numbers of targeted amplicons. Several modifications relative to our previously developed clean_tree tool (Ralf et al. 2015) were implemented to optimize the performance especially relevant for extremely large NGS data sets such as whole genomes. For instance, Yleaf extracts the Y-chromosomal reads prior to further processing and uses multithreading, a batch option is included too. Importantly, Yleaf provides drastically increased haplogroup resolution, that is, from 530 positions defining 432 NRY haplogroups with clean_tree (Ralf et al. 2015) to over 41,000 positions defining 5,353 haplogroups with Yleaf. For a detailed method description, see the Supplementary Material online.

To test for the performance of Yleaf in NRY-SNP calling, we compared variant calling with Yleaf, SAMtools/BCFtools (Li et al. 2009), and GATK (McKenna et al. 2010) in two whole genome data sets, one high and one low coverage, and achieved high concordance (for details, see supplementary table S4, Supplementary Material online). Moreover, to test the performance of Yleaf in NRY haplogroup inference, we analyzed 51 publically available NGS data sets produced from different starting materials of modern and ancient origin, with different library preparation methods, and different NGS platforms (table 1). In many cases, the NRY haplogroup resolution obtained with Yleaf was higher than obtained with other methods in the initial studies (table 1). The differences between previously reported and Yleaf-derived haplogroups of the nineteen 1000 Genomes Project samples are noteworthy as the haplogroups initially reported were inferred with the commercially available Yfull NextGen Sequence Interpretation, while Yleaf uses the public ISOGG tree. Although the results were generally compatible, in some cases Yleaf and in others Yfull provided the most detailed haplogroup. Compared with previously developed, noncommercial tools such as clean_tree (Ralf et al. 2015), AMY-tree (Van Geystelen et al. 2013), and yHaplo (Poznik 2016), Yleaf inferred Y-haplogroups with at least the same, and in the

Brief Communication

Table 1. Summary of NGS Data Sets Used for Automated NRY Haplogrouping with Yleaf.

Sample ID	No. of Detected SNPs	Average SNP Coverage	Predicted Haplogroup in Publication[a]	Yleaf Predicted Haplogroup	Most Downstream Predictive Marker(s)	Source
Modern DNA Whole Genome Sequencing						
HG00096_GBR	35,730	2.3×	R1b1a1a2a1a2c1a1f1a2[c]	R1b1a1a2a1a2c1a1f1a2	S691	1000 Genomes Project consortium (2015)
HG00190_FIN	31,814	1.9×	I1a1b1b1b1[b,c]	I1a1b1b1b1[b]	CTS1752 and CTS2783	1000 Genomes Project consortium (2015)
HG00329_FIN	36,255	2.5×	N1a1a1a1a2a1a2a[b]	N1a1a1a1a2a1a2a1[b]	CTS5057 and CTS6517	1000 Genomes Project consortium (2015)
HG00634_CHS	32,937	2.1×	O1b1a2a1	O1b1a2a1	F1759 and Z24393	1000 Genomes Project consortium (2015)
HG01097_PUR	36,118	2.4×	G2a2b2a1a1b1a1b	G2a2b2a1a1b1a1d[b]	CTS3664 and CTS10324	1000 Genomes Project consortium (2015)
HG01256_CLM	39,956	3.6×	J1[c]	J1	L255, M267, L321	1000 Genomes Project consortium (2015)
HG02645_GWD	40,911	4.6×	A1a[c]	A1a	V57 and V58	1000 Genomes Project consortium (2015)
HG03695_STU	38,934	3.3×	L1a2a1b[c]	L1a2a1b1[b]	Z34483, Z34486	1000 Genomes Project consortium (2015)
HG03705_PJL	35,934	2.4×	R1a1a1b2[c]	R1a1a1b2	F992	1000 Genomes Project consortium (2015)
HG03742_ITU	39,228	3.2×	NO[c]	NO	F549 and E482	1000 Genomes Project consortium (2015)
HG03745_STU	38,402	3.3×	H1a1a4b3b1a	H1a1a4b3b1a8[b]	Z34531 and Z34532	1000 Genomes Project consortium (2015)
HG03976_ITU	28,982	3.0×	J2b2[c]	J2b2	M241	1000 Genomes Project consortium (2015)
NA12154_CEU	37,798	3.1×	R1b1a1a2a1a2c1a1a1a1[c]	R1b1a1a2a1a2c1a1a1a1	DF23	1000 Genomes Project consortium (2015)
NA18486_YRI	37,642	2.9×	E1b1a1a1a2a1a3b1a2a1	E1b1a1a1a2a1a3b1a2a1b[b]	CTS11743 and CTS810	1000 Genomes Project consortium (2015)
NA18612_CHB	33,505	2.3×	C2c1a2b	C2c1a2b2[b]	PH404, Z31669	1000 Genomes Project consortium (2015)
NA18988_JPT	40,627	6.3×	D1b1a2b1a1a1[c]	D1b1a2b1a1a1b[b]	CTS2121 and CTS2897.	1000 Genomes Project consortium (2015)
NA19384_LWK	36,440	2.4×	B2b	B2b3[b]	CTS8235.2 and CTS8592	1000 Genomes Project consortium (2015)
NA19771_MXL	33,830	2.1×	Q1a2a1b1a2	Q1a2a1b1a2b[b]	K216.2, CTS11092, CTS649	1000 Genomes Project consortium (2015)
NA20348_ASW	36,294	2.4×	E2b2	E2b2b[b]	CTS2388 and CTS2641	1000 Genomes Project consortium (2015)
GM24149/HG003	41,520	26.3×	No data available	J1a2b3a1	L816	Zook et al. (2016)
Ancient DNA Whole Genome Sequencing						
WC1	38,772	6.0×	G2b	G2b2a	Z8022	Broushaki et al. (2016)
F38	26,181	1.7×	R1b1a1a2a	R1b1a1a2a2	PF7575 and M12149	Broushaki et al. (2016)
Bar31	31,896	2.5×	G2a2b2a1a1a1	G2a2b	F1733 and L32	Hofmanová et al. (2016)
Klei10	27,770	1.7×	G2a2a1a2	G2a2a1a2b	Z42731 and Z42572	Hofmanová et al. (2016)
3DRIF-16	12,503	1.2×	R1b1a1a2a1a1	R1b1a1a2a1a1c1a	S376	Martiniano et al. (2016)
3DRIF-26	18,628	1.3×	J2	J2b1	M205	Martiniano et al. (2016)
6DRIF-3	24,218	1.5×	R1b1a1a2a1a1	R1b1a1a2a1a1c1a1	DF98	Martiniano et al. (2016)
6DRIF-18	18,195	1.3×	R1b1a1a2a1a	R1b1a1a2a1a2c1a1i	FGC9661, FGC9658, FGC9655	Martiniano et al. (2016)
6DRIF-21	19,103	1.3×	R1b1a1a2a1a2c1b	R1b1a1a2a1a2c1b1	A94	Martiniano et al. (2016)
6DRIF-22	18,391	1.3×	R1b1a1a2a1a2b	R1b1a1a2a1a2b1d2a[b]	BY3497, BY3513	Martiniano et al. (2016)
6DRIF-23	12,397	1.2×	R1b1a1a2a1a	R1b1a1a2a1a	L52, PF6543, P310	Martiniano et al. (2016)
NO3423	17,721	1.3×	I1	I1a[b]	CTS9857	Martiniano et al. (2016)
S41	26,427	1.7×	D	D1a1a1a2	Z31603 and Z31605	Jeong et al. (2016)
C1	39,398	18.9×	O2a2b1a1	O2a2b1a1a6a	CTS5308	Jeong et al. (2016)
S10	33,403	2.1×	O2a2b1a1	O2a2b1a1a6a	CTS5308	Jeong et al. (2016)
S35	33,273	2.2×	O2a2b1a1	O2a2b1a1a6	CTS4658, CTS9332	Jeong et al. (2016)
SRR2544592	39,907	6.2×	E1b1	E1b1a2a[b]	Y17904, Y17905	Llorente et al. (2015)
PRJNA46213	37,652	30.1×	Q1a	Q1a1a2	Z36018 and Z36019	Rasmussen et al. (2010)
Modern DNA Whole Exome Sequencing						
HG00096_GBR	68,986	4.4×	R1b1a1a2a1a2c1a1f1a2[c]	R1b1a1a2a1a2c1a1f1	CTS6838	1000 Genomes Project consortium (2015)
HG00190_FIN	9,780	3.6×	I1a1b1b1b1[b,c]	I1a1b1b1b1[b]	CTS2783	1000 Genomes Project consortium (2015)
HG00329_FIN	20,088	2.7×	N1a1a1a1a2a1a2a[b]	N1a1a1a1a2a1a2a1[b]	CTS5057 and CTS6517	1000 Genomes Project consortium (2015)

(continued)

Table 1. Continued

Sample ID	No. of Detected SNPs	Average SNP Coverage	Predicted Haplogroup in Publication[a]	Yleaf Predicted Haplogroup	Most Downstream Predictive Marker(s)	Source
Modern DNA Targeted Capture Enrichment						
HGDP00001	5,293	14.1×	R1a1	R1a1a1	Page7, M417	Lippold et al. (2014)
HGDP00003	5,588	21.7×	L	L1a2a1b2[b]	SK1455, SK1454, Y18183	Lippold et al. (2014)
HGDP00005	5,901	17.2×	R2	R2a2b1b2b	SK2153	Lippold et al. (2014)
HGDP00007	5,748	24.6×	J2	J2a1b1	M92	Lippold et al. (2014)
HGDP00009	4,434	10.5×	J2	J2a1	L26, F4326	Lippold et al. (2014)
GRC12126890	38,455	93.6×	N1a2a1[b,c]	N1a2a1a[b]	F1988	Ilumäe et al. (2016)
GRC12126040	38,213	68.6×	N1b2[c]	N1b2	Z19753 and M1819	Ilumäe et al. (2016)
Modern DNA Amplicon Sequencing						
R1b1a2_1	1,006	540.3×	R1b1a1a2a1a2a1b1a1	R1b1a1a2a1a2a1b1a1	M167	Ralf et al. (2015)
R1b1a2_3	977	479.8×	R1b1a1a2a1a1c2b2b1a	R1b1a1a2a1a1c2b2b1a	Z326	Ralf et al. (2015)
R1b1a2_4	988	428.8×	R1b1a1a2a1a1c2b2	R1b1a1a2a1a1c2b2	S268, S379	Ralf et al. (2015)

[a]For comparative reasons, the nomenclature from the ISOGG was applied, using the most derived marker from the original publication.
[b]An approximate location in the ISOGG Y-tree, which may be relocated in future builds.
[c]Additional downstream markers that are included in the original publication, but (currently) not in the ISOGG tree.

majority of the samples with increased resolution (supplementary tables S2 and S3, Supplementary Material online). Another advantage of Yleaf compared with previous tools is that preprocessing of the NGS data is not needed as the Yleaf pipeline works with both raw and aligned sequencing data to produce the final haplogroup output files with a single command.

NGS data are not error-free; as a result, Yleaf can reveal Y-SNP calls that do not follow the phylogenetic pattern of the underlying NRY tree. However, this is where the vast amount of markers and the full consideration of the underlying NRY tree employed by Yleaf becomes evident. In cases of sequencing errors (likewise minor DNA contamination) the discordant Y-SNP calls will always be a minority, while the true haplogroup will be supported by the majority of the calls. In addition, the discordant Y-SNP calls will not be supported by their own upstream and equivalent markers and therefore are easily interpreted as the result of a sequencing error, minor contamination, or may in part reflect private mutations. Thus, Yleaf allows for the correct haplogroup assignment despite potential sequencing errors, minor contamination, and private mutations. In the WGS data produced from ancient samples, the frequency of discordant SNP calls was considerably higher compared with WGS from modern DNA samples (see Supplementary Material online). This can be explained by the increased sequencing errors due to poor DNA quality and/or increased risk of contamination. Yleaf also revealed some self-contradictory results that may indicate the need for topological revisions in the underlying NRY tree such as in sample HG01097 (see Supplementary Material online). For a detailed result description, see the Supplementary Material online.

Yleaf with an installation guide is publically available at https://www.erasmusmc.nl/genetic_identification/resources/, last accessed March 28, 2018 as well as the output files from the 51 analyzed samples. Since more NRY NGS data become available constantly and ISOGG is updating the NRY-tree regularly, Yleaf will be regularly updated.

In conclusion, we introduce and make publically available Yleaf, an easy-to-use, highly flexible software tool for accurate, high-resolution haplogroup inference from Y-chromosome NGS data of all types that is independent of sample preparation and sequencing technology and outperforms previously developed tools. We envision that Yleaf will serve the community that uses NRY variation from NGS data for various research and application purposes.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

This work was supported by Erasmus MC University Medical Center Rotterdam. K.Z. was supported by the China Scholarship Council to carry out his PhD studies at Erasmus MC.

References

Broushaki F, Thomas MG, Link V, López S, van Dorp L, Kirsanow K, Hofmanová Z, Diekmann Y, Cassidy LM, Díez-Del-Molino D, et al. 2016. Early Neolithic genomes from the eastern Fertile Crescent. *Science* 353(6298):499–503.

Hallast P, Batini C, Zadik D, Maisano Delser P, Wetton JH, Arroyo-Pardo E, Cavalleri GL, de Knijff P, Destro Bisol G, Dupuy BM, et al. 2015. The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades. *Mol Biol Evol.* 32(3):661–673.

Hofmanová Z, Kreutzer S, Hellenthal G, Sell C, Diekmann Y, Díez-del-Molino D, van Dorp L, López S, Kousathanas A, Link V, et al. 2016. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc Natl Acad Sci U S A.* 113(25):6886–6891.

Ilumäe A-M, Reidla M, Chukhryaeva M, Järve M, Post H, Karmin M, Saag L, Agdzhoyan A, Kushniarevich A, Litvinov S, et al. 2016. Human Y chromosome haplogroup N: a non-trivial time-resolved phylogeography that cuts across language families. *Am J Hum Genet.* 99(1):163–173.

Jeong C, Ozga AT, Witonsky DB, Malmström H, Edlund H, Hofman CA, Hagan RW, Jakobsson M, Lewis CM, Aldenderfer MS, et al. 2016. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. *Proc Natl Acad Sci U S A.* 113(27):7485–7490.

Jobling MA, Tyler-Smith C. 2017. Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet.* 18(8):485–497.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* 25(16):2078–2079.

Lippold S, Xu H, Ko A, Li M, Renaud G, Butthof A, Schröder R, Stoneking M. 2014. Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investig Genet.* 5:13.

Llorente MG, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, Curtis MC, Stock JT, Coltorti M, Pieruccini P, et al. 2015. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. *Science* 350(6262):820–822.

Martiniano R, Caffell A, Holst M, Hunter-Mann K, Montgomery J, Müldner G, McLaughlin RL, Teasdale MD, van Rheenen W, Veldink JH, et al. 2016. Genomic signals of migration and continuity in Britain before the Anglo-Saxons. *Nat Commun.* 7:10326.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 20(9):1297–1303.

Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, et al. 2016. Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nat Genet.* 48(6):593–599.

Poznik GD. 2016. Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men [cited 2017 Nov 23]. Available from: https://www.biorxiv.org/content/early/2016/11/19/088716.

Ralf A, Oven M, Zhong K, Kayser M. 2015. Simultaneous analysis of hundreds of Y-chromosomal SNPs for high-resolution paternal lineage classification using targeted semiconductor sequencing. *Hum Mutat.* 36(1):151–159.

Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, Metspalu M, Metspalu E, Kivisild T, Gupta R, et al. 2010. Ancient human genome sequence of an extinct Palaeo-Eskimo. *Nature* 463(7282):757.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. *Nature* 526:68–74.

Van Geystelen A, Decorte R, Larmuseau M. 2013. AMY-tree: an algorithm to use whole genome SNP calling for Y chromosomal phylogenetic applications. *BMC Genomics* 14(1):101.

Zook JM, Catoe D, McDaniel J, Vang L, Spies N, Sidow A, Weng Z, Liu Y, Mason CE, Alexander N, et al. 2016. Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci Data* 3:160025.