Forensic Science International: Genetics Supplement Series 8 (2022) 97–98

Contents lists available at ScienceDirect

Forensic Science International: Genetics Supplement Series

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

Targeted Y chromosome capture enrichment in admixed South American samples with haplogroup Q

Zehra Köksal [a,*], Germán Burgos [b,c], Elizeu Carvalho [d], Humberto Ossa [e,f], María Laura Parolin [g], Alfredo Quiroz [h], Ulises Toscanini [i], Carlos Vullo [j], Claus Børsting [a], Leonor Gusmão [d], Vania Pereira [a]

[a] Section of Forensic Genetics, Department of Forensic Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark
[b] Escuela de Medicina, Facultad de Ciencias de la Salud, Universidad de Las Américas (UDLA), Quito, Ecuador
[c] Grupo de Medicina Xenómica, Universidad de Santiago de Compostela, Santiago de Compostela, Spain
[d] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Rio de Janeiro, Brazil
[e] Laboratório de Genética y Biología Molecular, Bogotá, Colombia
[f] Pontificia Universidad Javeriana, Facultad de Ciencias, Bogotá, Colombia
[g] Instituto de Diversidad y Evolución Austral (IDEAus), Centro Nacional Patagónico, CONICET, Puerto Madryn, Argentina
[h] Instituto de Previsión Social, Asunción 100153, Paraguay
[i] Primer Centro Argentino de Inmunogenética (PRICAI), Fundación Favaloro, Buenos Aires, Argentina
[j] DNA Forensic Laboratory, Argentine Forensic Anthropology Team (EAAF), Córdoba, Argentina

ARTICLE INFO

Keywords:
Y chromosome
Haplogroup Q
Targeted capture
MPS
NGS

ABSTRACT

Y haplogroups, defined by Y-SNPs, allow the reconstruction of the human Y chromosome genealogy. Recently, MPS based panels were introduced in the forensic genetics community for Y-SNP typing and identification of a broad range of haplogroups. The panels are based on an amplicon strategy and allow the detection of up to 15,600 Y-SNPs. The panels target up to 210,000 bps, which should be compared to the overall 8.9 Mbps comprising the unique regions of the non-recombining portion of the Y chromosome (NRY). We present an alternative approach of sequencing unique regions within the NRY using target enrichment probes and hybridization capture. A total of 359,954 probes were designed using the SureDesign software, representing 7.5 Mbps of the NRY. Library preparation and capture were performed using the Agilent SureSelect XT HS2 Target Enrichment method and sequencing was performed in a NovaSeq 6000 System. Besides individual barcodes, the method also included unique molecular barcodes for additional quality screening. The method was tested on admixed South Americans that carry a Y chromosome of haplogroup Q. We successfully identified novel variation that could potentially help refining haplogroup Q phylogeny.

1. Introduction

Y-SNPs were accumulated over generations and define Y-chromosomal haplogroups, which contain information that can be used for evolutionary, population and forensic genetics [1]. To this date, more than 70,000 Y-SNPs distributed across the non-recombining portion of the Y chromosome (NRY) are reported in public databases (ISOGG Y-DNA Haplogroup Trees 2019–2020), and 20 main haplogroups have been described.

Current methods target up to 9014 phylogenetic informative Y-SNPs using amplicon-based Massively Parallel Sequencing panels [2,3]. These panels target up to 209,248 bps, which is only a small portion of the NRY. Additionally, an uneven representation of Y-SNPs from certain Y-chromosomal haplogroups can be observed: Haplogroup R is a well-researched haplogroup making up 17–24 % (153 [2] and 2142 [3] Y-SNPs) of the targeted SNPs. An underrepresentation of some haplogroups, like haplogroup Q, leads to biases in the haplogroup estimations. An unbiased representation of the Y-SNPs can be achieved by sequencing the entire NRY [4–7]. Whole genome sequencing could be a possible approach, but it produces an extremely high amount of data that are not relevant for the study of the Y chromosome.

In this work we present a hybridization-based capture protocol for

* Correspondence to: Frederik V's vej 11, 2100 Copenhagen, Denmark.
E-mail address: Zehra.koksal@sund.ku.dk (Z. Köksal).

https://doi.org/10.1016/j.fsigss.2022.09.034
Received 6 September 2022; Accepted 29 September 2022
Available online 30 September 2022

the analysis of admixed South Americans of haplogroup Q. The prepared libraries were captured using 359,954 RNA probes covering 7.5 Mbps of unique regions within the NRY. The final captured libraries were amplified and sequenced on Illumina's NovaSeq 6000 System (San Diego, CA, USA).

2. Material and methods

A total of 16 admixed South Americans were selected. Following the manufacturer's protocol, 11–15 ng of DNA input was fragmented using the Covaris S220 Focused Ultrasonicator (Covaris LLC., MA, USA). Libraries were generated using Agilent's SureSelect XT HS2 Target Enrichment for the Illumina Paired-End Sequencing Library kit (Agilent Technologies, CA, USA) according to the manufacturer's protocol.

The regions of interest were captured by adding RNA probes complementary to the unique regions within the Y chromosome [4]. Custom-made probes were designed using Agilent's SureDesign SureSelectDNA software tool, resulting in 359,954 RNA probes covering 7.5 Mbps of the unique regions within the NRY. Combining the RNA probes and libraries resulted in hybridization of complementary probes and library molecules, which were captured, purified and amplified. All 16 libraries pooled equimolarly to 2 nM were paired-end sequenced using the NovaSeq 6000 SP Reagent Kit v1.5 (300 cycles) on the Illumina NovaSeq 6000 System (Illumina, CA, USA). For the data analysis, an in-house generated pipeline was applied that included the removal and annotation of molecular barcodes (MBC), alignment to the reference genome GRCh37, sorting of the BAM files, and deduplication using the MBCs (duplex consensus mode). Variants were called within the target regions with a minimum base quality of 25.

3. Results and discussion

Among the 16 samples, the fraction of on-target sequenced bases ranged from 19 % to 53 % (average: 43 ± 9 %) (Fig. 1). Bigger fractions of 36–80 % (average 68 ± 13 %) were assigned to the whole Y chromosome, which included near-target regions that were sequenced due to the random fragmentation of the sonication. Chromosome 1 was the second most represented chromosome, comprising no more than 6 % of all sequenced bases.

After deduplication of molecular barcodes, the total number of sequenced bases ranged from 147 to 217 Mb per sample (average: 173 ± 38 Mb). Within the targeted regions, 4.84 Mb were common to 13 samples with minimum coverage of 5 reads. This equals 65% of the targeted regions.

A total of 1921 variants (1703 SNPs and 218 indels) were found within the targeted regions in the 16 samples. Of these, 1179 variants were found in only one individual. A total of 545 and 225 SNPs had previously been reported in the dbSNP database and for the ISOGG Y-DNA Haplogroup Tree 2019–2020, respectively. In the ISOGG tree, 190 SNPs were associated with haplogroup Q, and the remaining 35 SNPs to other haplogroups. Fifteen of the indels had previously been reported in the dbSNP database, and one haplogroup Q indel was reported in the ISOGG Y-DNA Haplogroup Trees 2019–2020. The remaining variants (933 SNPs and 202 indels) were, to our knowledge, novel and can potentially entail variation, that may help to further resolve deeper branches of haplogroup Q.

4. Conclusion

In the present study, we aimed to explore an alternative approach to sequencing a wide informative region of the Y chromosome to find (novel) variation. The current method offers high specificity for the Y chromosome (68 ± 13 %) and the selected target regions (43 ± 9 %). After stringent quality controls, 4.84 Mbps were covered in > 80 % of the samples with minimum 5× read depth. Within the target regions of all 16 samples, we found 1921 variants including 933 SNPs and 202 indels that were not previously annotated and harbor the potential to differentiate sub-haplogroups within haplogroup Q.

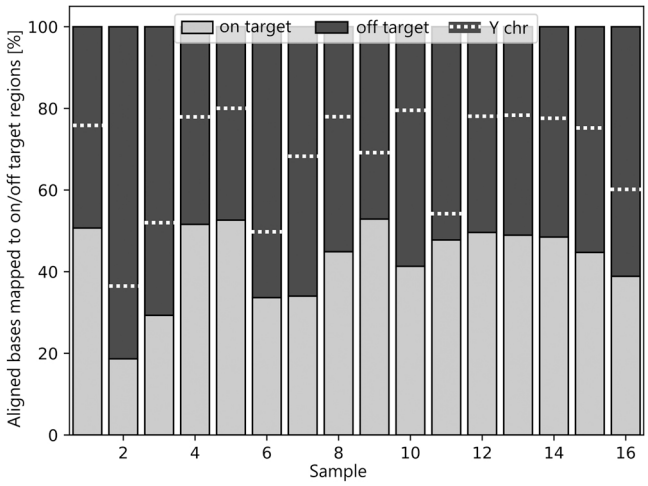

Fig. 1. Percentage of aligned bases that mapped to on-target regions (light gray) and off-target regions (dark gray) per sample. Fraction of bases assigned to the Y chromosome was marked with white dashed line.

Conflict of interest

None.

Acknowledgments

The authors would like to thank all the donors for volunteering to provide DNA samples.

References

[1] P.A. Underhill, et al., Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography, Genome Res, 7 (10) (1997) 996–1005, https://doi.org/10.1101/gr.7.10.996.
[2] A. Ralf, et al., Forensic Y-SNP analysis beyond SNaPshot: high-resolution Y-chromosomal haplogrouping from low quality and quantity DNA using Ion AmpliSeq and targeted massively parallel sequencing, Forensic Sci. Int. Genet. 41 (2019) 93–106, https://doi.org/10.1016/j.fsigen.2019.04.001.
[3] S. Claerhout, P. Verstraete, L. Warnez, S. Vanpaemel, M. Larmuseau, R. Decorte, CSYseq: the first Y-chromosome sequencing tool typing a large number of Y-SNPs and Y-STRs to unravel worldwide human population genetics, PLOS Genet. 17 (9) (2021), e1009758, https://doi.org/10.1371/journal.pgen.1009758.
[4] W. Wei, et al., A calibrated human Y-chromosomal phylogeny based on resequencing, Genome Res. 23 (2) (2013) 388–395, https://doi.org/10.1101/gr.143198.112.
[5] P. Hallast, et al., The Y-chromosome tree bursts into leaf: 13,000 high-confidence SNPs covering the majority of known clades, Mol. Biol. Evol. 32 (3) (2015) 661–673, https://doi.org/10.1093/molbev/msu327.
[6] D.I. Cruz-Dávalos, et al., In-solution Y-chromosome capture-enrichment on ancient DNA libraries, BMC Genom. 19 (1) (2018) 608, https://doi.org/10.1186/s12864-018-4945-x.
[7] A.B. Rohrlach, et al., Using Y-chromosome capture enrichment to resolve haplogroup H2 shows new evidence for a two-path Neolithic expansion to Western Europe, Sci. Rep. 11 (1) (2021) 15005, https://doi.org/10.1038/s41598-021-94491-z.