J. Biol. Res. 2012 n.1 - Vol. LXXXV - Rubbettino - Soveria Mannelli

Fragment Classify Tool on Trial: West Papua Sample

F.M. Calabrese[1], F. Rubino[1], M. Tommaseo-Ponzetta[2], M. Attimonelli[1]

[1] Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche Università di Bari, Bari, 70126, Italy.
E-mail: franccalabrese@libero.it
[2] Dipartimento di Biologia, Università di Bari, Bari, 70126, Italy

KEY WORDS: mtDNA, haplogroups, HmtDB database, Papua.

Introduction

Human mitochondrial haplogroups classification has become a recurrent theme in the last decades. Mitochondrial DNA (mtDNA) haplogroups number and deepness is continuously increasing, thanks to the growth in type and quantity of the molecular data that become available from mitochondrial population genetics and clinical studies. This will allow the design of a complete pathway, linking the history of their spread with specific sub-continent areas variability. Although Next Generation Sequencing (NGS) technology facilitates the sequencing of the entire mtDNA molecule, the sequencing of the HVS1 and HVS2 regions located within the D-loop, the major mtDNA regulatory region, is still a practice in use. The production of a great quantity of complete human mitochondrial genomes has required the availability of an *ad hoc* organized genomic resource. HmtDB (Rubino *et al.*, 2011) is an open resource created in the Bioinformatics Lab in Bari, to support population genetics and mitochondrial disease studies. The database, which hosts more than 8000 human mitochondrial genome sequences annotated with population and variability data, is continuously updated. The comparison of each complete mtDNA sequence with the revised Cambridge Reference Sequence (rCRS, GenBank Accession NC_012920.1) allows to identify variant sites and hence to assign the sequence to the most appropriate haplogroup. Moreover, the occurrence of individual-specific, or private, mutations may lead either to define a new sub-haplogroup or to highlight disease association or last but not least to evaluate the quality of the sequence. *Phylotree* is a worldwide recognized reference system that gathers these data into a complete and updated human mtDNA haplogroup classification tree [http://www.phylotree.org] (van Oven *et al.*, 2009). The haplogroup prediction can be performed by applying two different tools available within HmtDB: the "classify your genome" and "fragment classify". The latter has been implemented to predict haplogroups even on incomplete mtDNA sequences. Here we report the application of the "fragment classify" tool to the gathering of more than 500 sequences belonging to different Papua New Guinea groups.

Materials and Methods

Haplogroup classification has been carried out by using the "fragment classify" tool available in HmtDB database (http://www.hmtdb.uniba.it:8080/hmdb/). The software accepts as input both single fasta and multifasta files. It selects for each haplogroup the variant sites annotated in Phylotree (N_{ph}), among the total number of SNPs located in the fragment region. Furthermore, the tool reports the number of SNPs defining the haplogroup in the whole genome (N_{ph_tot}) and the N_{ph_tot} subset expected in the fragment N_{ph_exp}. The assigned haplogroup is the one whose fraction of N_{ph} over the total number of the haplogroup-defining expected sites (N_{ph_exp}) is highest. The higher is the ratio N_{ph_exp}/N_{ph_tot}, the more robust is the assignment. The number of N_{ph} sites defining the haplogroups is strictly related to the evolutionary distance between rCRS and the considered genomes. Hence, the higher is the distance, the lower could be the reliability of the obtained prediction as much as shorter is the fragment. Since partial genomes are also annotated in HmtDB, the Author/Fragment_Classifier Predicted Haplogroup code has been introduced within the Genome Card, in order to correct the haplogroup prediction, which may not be obtained with the entire genome classifier. Next to the haplogroup assigned by the author, the best haplogroup prediction, selected according to a sorting algorithm, upon application of the fragment classifier is indicated. The tool is written in Python and can be downloaded from HmtDB site, as a package to be locally implemented.

Results and Discussion

Although sequencing of the entire mitochondrial genome is becoming a feasible approach, frequently researchers prefer to focus on short fragments of the genome. In these circumstances, due to the scattering of the defining sites of any specific haplogroup along the entire mtDNA molecules, the sensibility of the classification tool available in HmtDB within the option "Classify your genome" becomes

increasingly lower as the decrease in the fragment length. The fragment-haplo-classifier tool, inclusive of the sorting algorithm, overcomes this problem as confirmed by the results reported below.

An example from West Papua - 555 partial mtDNA sequences were considered, representative of 30 populations of Papua, the Indonesian part of New Guinea Island. 265 sequences belong to populations of the Bird's Head Peninsula and north-western surrounding regions, while 290 are representative of human groups living in the central and south-western parts of the region. The sampled populations speak both Austronesian and Non-Austronesian (or Papuan) languages. The haplogroups of the above sequences, part of which already published (Tommaseo-Ponzetta *et al.*, 2002, 2007; Cascione *et al.* 2008) were assigned or reassigned using the Fragment classify tool. Since haplogroups classification is continuously updated, this follow up resulted in a more reliable haplogroup attribution. Q is confirmed as the most frequent haplogroup in Papua, nesting 66% of maternal lineages, whereas the frequency of P reaches 20% and is mostly found among central highlands pygmoid groups (Fig. 1). Other haplogroups, belonging to different branches of B, E, D, G or F haplogroups are also present, especially along the Bird's Head coastal areas, confirming the continuous contacts with the surrounding archipelagos.

References

Cascione I., Attimonelli M., Syukriani Y., Noer S.A., Marzuki S., Tommaseo Ponzetta M., 2008. La variabilità mitocondriale delle popolazioni del Bird's Head. *Int. J. Anthrop.*, N.S.: 38-42.

Rubino F., Piredda R., Calabrese F.M., Simone D., Lang M., Calabrese C., Petruzzella V., Tommaseo-Ponzetta M., Gasparre G., Attimonelli M., 2011. HmtDB, a genomic resource for mitochondrion-based human variability studies. *Nucleic Acids Res.*, 2012,40(D1): D1150-D1159.

Tommaseo Ponzetta M., De Robertis M., Tanzariello F., Attimonelli M., Saccone C., 2002. Mitochondrial DNA variability in W. Papua populations. *Am. J. Phys. Anthrop.*, 117(1): 49-67.

Tommaseo-Ponzetta M., Cascione I., Attimonelli M., Sudoyo H., Marzuki S., 2007. Mitochondrial DNA M and N haplogroups in West New Guinea populations. *Recent Advances on Southeast Asian Paleoanthropology and Archaeology* (ed. E. Indriati), Gadjah Mada University, Yogyakarta: 207-215.

Van Oven M., Kayser M., 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.*, 2009(2): E386-394.

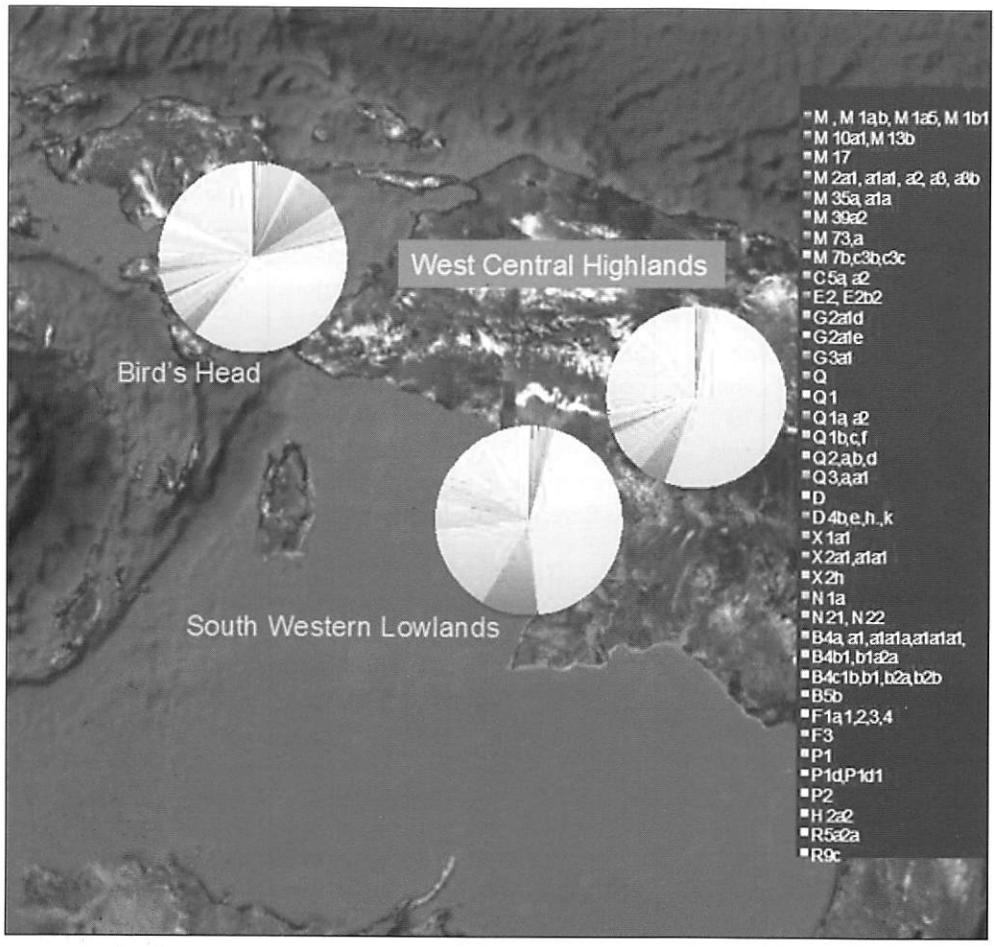

Fig. 1. *MtDNA haplogroups's frequency distribution in West Papua.*