H+, or how to build a perfect human.

Author: Kristina Zheltova

Abstract

In this work we worked on analysis raw 23andMe data, trying to obtain such information like maternal and paternal haplogroups, eye and skin colors, clinically relevant data from SNPs.

Introduction

Each of us carries a large number of genetic variations. Using a genome sequence obtained by an NGS or just a collection of SNP we could try to predict likelihood of having some phenotypic trait, or likelihood of disease.

Also, there is CRISPR/Cas9 – a specific, efficient and versatile gene-editing technology we can harness to modify, delete or correct precise regions of our DNA [1].

Methods

Raw 23andMe data were taken from here.

For analysis we converted 23andMe's raw data into standard .vcf format. We used plink [2]. Mthap [3] was used to identify all SNPs that distinguish the haplogroup. To identify paternal (Y chromosome) haplogroups. we used MorleyDNA Y-SNP Subclade Predictor [4].

For getting annotation of all SNPs we used online Variant Effect Predictor – GRCh37 [5].

Results

After removing all SNPs corresponding to deletions and insertions to make the file compatible with annotation tools, we established probable ethnicity of given subject by identifying maternal (Figure 1) and paternal haplogroups (Figure 2).

We also tried to determine this person’s sex and eye and skins colors based on approach with 8 SNPs using [6]. So, eye color detected as brown, skin color as Non-dark skin color (ie, light or medium), sex as male.

Annotation of the obtaining SNPs and extracting all clinically relevant SNPs were done (Figure 3). For example, rs1024611 17:32579788-32579788 G shows susceptibility to coronary artery disease, development of, in HIV. Such information obtained with ClinVar [7].

Markers found (shown as differences to rCRS):

HVR2: 152C 263G
CR: 750G 1438G 4769G 8860G
HVR1:

IMPORTANT NOTE: The above marker list is almost certainly incomplete due to limitations of genotyping technology and is not comparable to mtDNA sequencing results. It should not be used with services or tools that expect sequencing results, such as mitosearch.

Best mtDNA Haplogroup Matches:

1) H(T152C)

Defining Markers for haplogroup H(T152C):
HVR2: 152C 263G
CR: 750G 1438G 4769G 8860G 15326G
HVR1:

Marker path from rCRS to haplogroup H(T152C):
H2a2a1(rCRS) ⇨ 263G ⇨ H2a2a ⇨ 8860G 15326G ⇨ H2a2 ⇨ 750G ⇨ H2a ⇨ 4769G ⇨ H2 ⇨ 1438G ⇨ H ⇨ 152C ⇨ H(T152C)

Imperfect Match. Your results contained differences with this haplogroup:
Matches(6): 152C 263G 750G 1438G 4769G 8860G
Untested(1): 15326

2) H1(T152C)

Defining Markers for haplogroup H1(T152C):
HVR2: 152C 263G
CR: 750G 1438G 3010A 4769G 8860G 15326G
HVR1:

Marker path from rCRS to haplogroup H1(T152C):
H2a2a1(rCRS) ⇨ 263G ⇨ H2a2a ⇨ 8860G 15326G ⇨ H2a2 ⇨ 750G ⇨ H2a ⇨ 4769G ⇨ H2 ⇨ 1438G ⇨ H ⇨ 3010A ⇨ H1 ⇨ 152C ⇨ H1(T152C)

Imperfect Match. Your results contained differences with this haplogroup:
Matches(6): 152C 263G 750G 1438G 4769G 8860G
No-Calls(1): 3010A
Untested(1): 15326

Figure 1 – mthap results fragment.

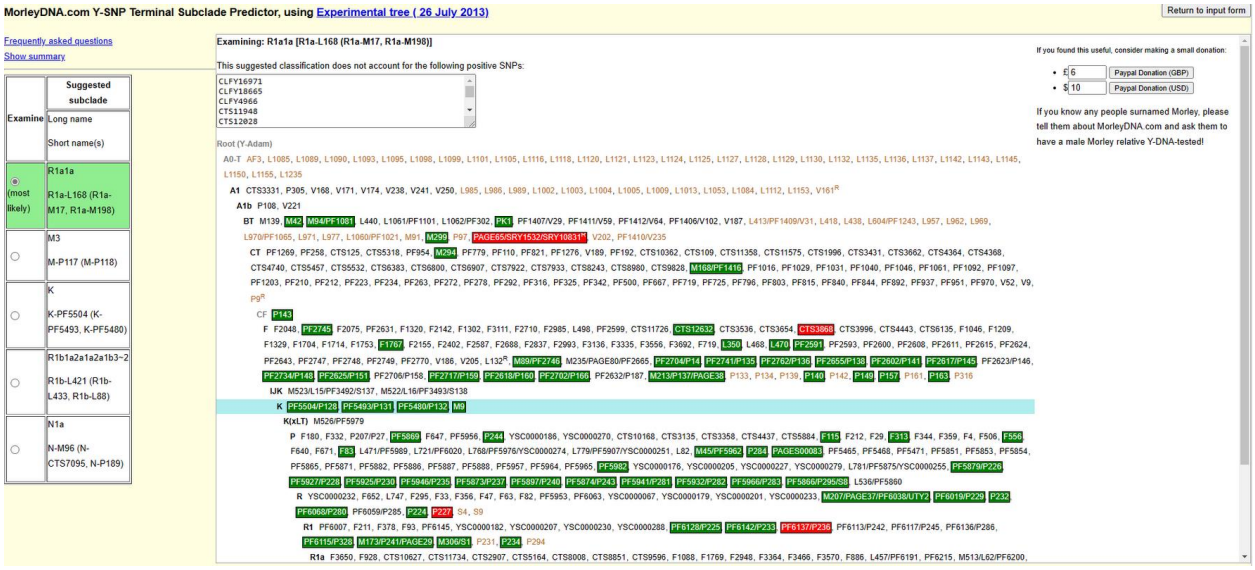

Figure 2 – MorleyDNA results fragment.

```
i3000469        2:138759649-138759649   T
i6007787        2:234183368-234183368   G
i6058143        1:161479745-161479745   G
i6059141        8:133909974-133909974   G
rs1024611       17:32579788-32579788    G
rs1049296       3:133494354-133494354   T
rs10757274      9:22096055-22096055     G
rs1169288       12:121416650-121416650  C
rs12150220      17:5485367-5485367      T
rs13266634      8:118184783-118184783   T
rs1801197       7:93055753-93055753     G
rs1801274       1:161479745-161479745   G
rs1801275       16:27374400-27374400    G
rs1801394       5:7870973-7870973       G
rs1801968       9:132580901-132580901   G
rs2004640       7:128578301-128578301   T
rs2073658       1:161010762-161010762   T
rs2184026       9:101304348-101304348   T
rs2239704       6:31540141-31540141     C
rs2241880       2:234183368-234183368   G
rs2281845       1:201081943-201081943   T
rs231775        2:204732714-204732714   G
rs4402960       3:185511687-185511687   T
rs4880  6:160113872-160113872   G
rs4961  4:2906707-2906707       T
rs4977574       9:22098574-22098574     G
rs5174  1:53712727-53712727     T
rs5186  3:148459988-148459988   C
rs61747071      16:53720436-53720436    T
rs6265  11:27679916-27679916    T
rs6280  3:113890815-113890815   T
rs6504649       17:48437456-48437456    G
rs699   1:230845794-230845794   G
rs763110        1:172627498-172627498   T
rs7794745       7:146489606-146489606   T
rs909253        6:31540313-31540313     G
```

Figure 3 - Extracting clinically relevant SNPs.

References

1. CRISPR/Cas9
2. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559-75. doi: 10.1086/519795. Epub 2007 Jul 25. PMID: 17701901; PMCID: PMC1950838.
3. Mthap
4. MorleyDNA Y-SNP Subclade Predictor
5. GRCh37
6. Hart KL, Kimura SL, Mushailov V, Budimlija ZM, Prinz M, Wurmbach E. Improved eye- and skin-color prediction based on 8 SNPs. Croat Med J. 2013 Jun;54(3):248-56. doi: 10.3325/cmj.2013.54.248. PMID: 23771755; PMCID: PMC3694299.
7. ClinVar