Haplogroup-R.org uploaded files

Haplogroup-R.org meta-data

FTDNA public data
FGC lobSTR data

Haplogroup project data
Literature data
YSeq, etc.

Kits

Sequence ID	16b integer
VCF filename	256B string
BED filename	256B string
Company	64B string
Test Type	64B string
Kit Number	64B string
Date of upload	16b integer
Person	**16b integer**
STRs	512*8b integer array

65536 kits at 1222B/entry = 76 MB

People

Person	16b integer
Kit sequence Ids	**4*16b integer array**
NGS tested?	logical
MDKA surname	64B string
Latitude	16b float
Longitude	16b float
MDKA descriptor	256B string
MDKA date	16b float
- " - uncertainty	16b float
MRKH Node ID	**16b integer**
Shared coverage	32b integer
" within age BED	32b integer
Total coverage	32b integer
*Most-recent known haplogroup	

65536 kits at 352.25B/entry = 22 MB

Calls
2D array with *n* **people** and *m* **variants**

Elements:

Positive	logical
Callable*	logical
Read quality	8b integer
Mapping quality	8b integer
Read depth	8b integer
Derived reads	8b integer
Ancestral reads	8b integer

*E.g. Presumed positives would be positive=TRUE callable=FALSE

*524288 variants * 65536 people * 5.25 bytes = 168 GB*

Variants

Sequence ID	32b integer
Primary name	16B string
Alt names	3*16B string array
GRCh37 start	32b integer
GRCh38 start	32b integer
SNP/MNP*	logical
Ref^	32b base 4 integer
Anc^	32b base 4 integer
Der^	32b base 4 integer
Source	16B string
Within age BED	logical

*MNP includes indels
^Could be 16B text strings

*524288 variants * 62.25 bytes/entry = 31 MB*

Variant database
Reference haplotree
BED file for age analysis

Tree

Node ID	16b integer
Parent node ID	16b integer
Clade name	16B string
Variants	**64*16b integer array**
Quality scores*	64*logical array
Age analysis flags	64*logical array
Child clades	64*16b integer array
Ancestral STRs	512*8b integer array
Origin	2*16b float array
Shared coverage	32b integer
" within age BED	32b integer
Oldest MRCA	16b signed integer
Oldest MRCA kit	16b integer
SNP age	16b integer
SNP age 95% c.i.	2*16b integer array
SNP age PDF	1024*16b float array
SNP parent PDF	1024*16b float array
STR age	16b integer
STR age 95% c.i.	2*16b integer array
STR age PDF	1024*16b float array
STR parent PDF	1024*16b float array
Combined age	16b integer
Combined 95% c.i.	2*16b integer array
Combined PDF	1024*16b float array

*Quality scores may denote precise/approximate placements or may be floats to contain more information.

65536 nodes at 11106 bytes/entry = 694 MB

Current database size is about 4000 NGS kits, 4000 NGS-tested people and 60000 variants, giving ~1.2 GB for the **Calls** structure.

Calls only needs to include the subset of people with NGS tests (including archive/literature DNA) and YSeq tests which split up clades.