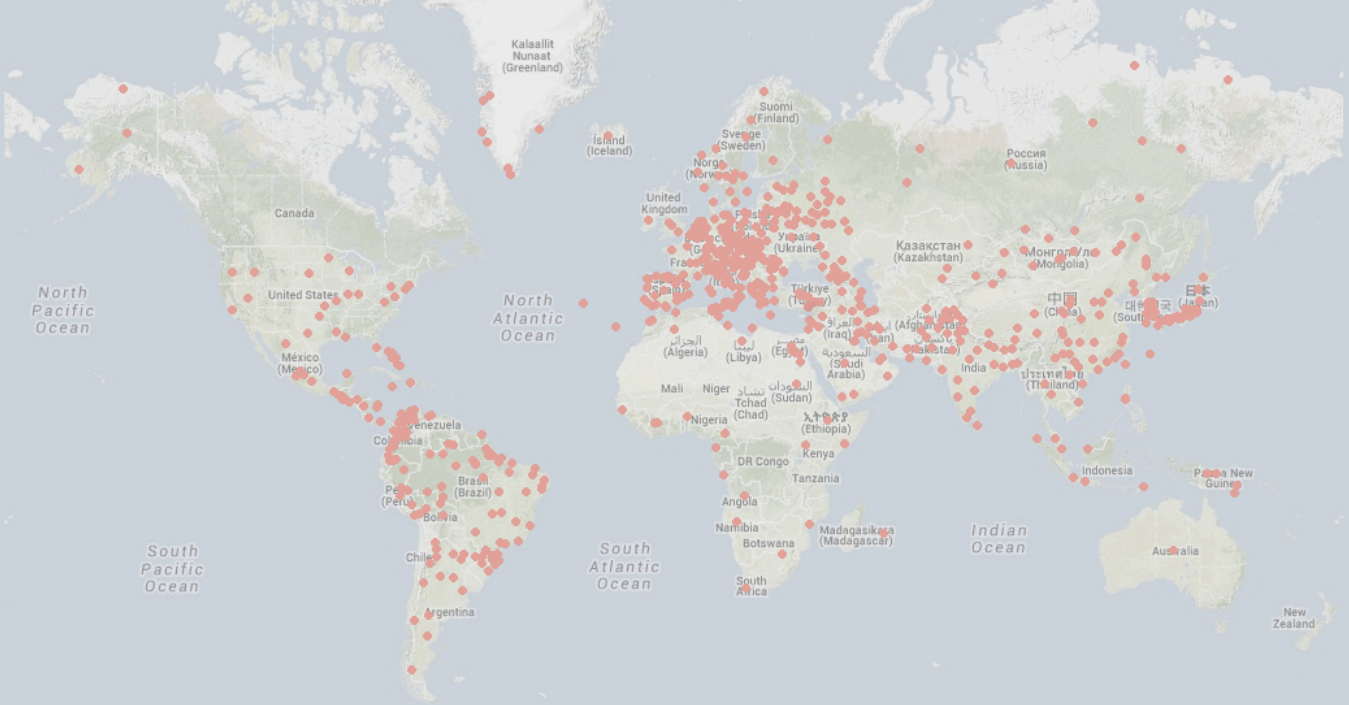


The Y-Chromosome Haplotype Reference Database

– Directions for Use –

Copyright and curation: Sascha Willuweit & Lutz Roewer
Institute of Legal Medicine and Forensic Sciences, Charité – Universitätsmedizin Berlin

Endorsed by International Society of Forensic Genetics (ISFG)
Supported by Life Technologies and Promega

Revision 44 • September 1st, 2013

Corresponding address
Sascha Willuweit & Lutz Roewer
Charité – Universitätsmedizin Berlin
Institute of Legal Medicine and Forensic Sciences
Department of Forensic Genetics
Augustenburger Platz 1
D-13353 Berlin
Germany

E-Mail
Sascha Willuweit: sascha DOT willuweit AT charite DOT de
Lutz Roewer: lutz DOT roewer AT charite DOT de

Contents

1 Introduction

The YHRD[1] (see Figure 1) aims to help with the interpretation of results from comparisons of evidentiary samples typed with Y-STRs and reference samples and to formulate conclusions. Since Y-STRs are located on the non-recombining part of the Y chromosome the profile generated by Y-STR analysis should be considered as one trait coded by one locus (a Haplotype). Consequently, the YHRD provides Haplotype frequencies (>1 locus typed per sample) for common formats consisting of 9-23 loci (see Table 1).

Name	#	Loci	Commercial Kits
Minimal	9	DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 and DYS385a/b	Applied Biosystems AmpFℓSTR® Yfiler®[2], Biotype Mentype® Argus Y-MH^{QS}[3], Promega PowerPlex® Y[4], Promega PowerPlex® Y23[5]
SWGDAM	11	Minimal + DYS438 and DYS439	Applied Biosystems AmpFℓSTR® Yfiler®, Promega PowerPlex® Y, Promega PowerPlex® Y23
PowerPlex Y	12	SWGDAM + DYS437	Applied Biosystems AmpFℓSTR® Yfiler®, Promega PowerPlex® Y, Promega PowerPlex® Y23
YFiler	17	PowerPlex + DYS448, DYS456, DYS458, DYS635 and YGATAH4	Applied Biosystems AmpFℓSTR® Yfiler®, Promega PowerPlex® Y23
PowerPlex Y23	23	YFiler + DYS576, DYS481, DYS549, DYS533, DYS570 and DYS643	Promega PowerPlex® Y23

Table 1: Available loci sets

Since a strong substructure between and within continents exists for Y-STR Haplotypes, the database must reflect the enrichment, scarcity or admixture of Haplotypes in certain geographical regions or continents. In contrast, different populations which share a common ancestry have shown to be homogeneous. Such population samples with none or small genetic distance can be joined to so-called Metapopulations (e.g. the Western Europeans). Dependent on your choice the database reports matches for the full database, a selected Metapopulation, a national database or a certain population sample. The database is fed by quality-assessed and certified forensic and other labs from all around the world. Thus, the YHRD is a meta-database aiming to implement all national database projects. The database is explained in detail in Willuweit et al. (2007) and Roewer (2009).

[1] See http://www.yhrd.org
[2] See http://www.lifetechnologies.com/order/catalog/product/4359513
[3] See http://www.biotype.de/en/products/forensics/mentyper-argus-y-mhqs.html
[4] See http://www.promega.com/applications/hmnid/profiles/powerplexy.htm
[5] See http://www.promega.com/products/pm/genetic-identity/powerplex-y23/

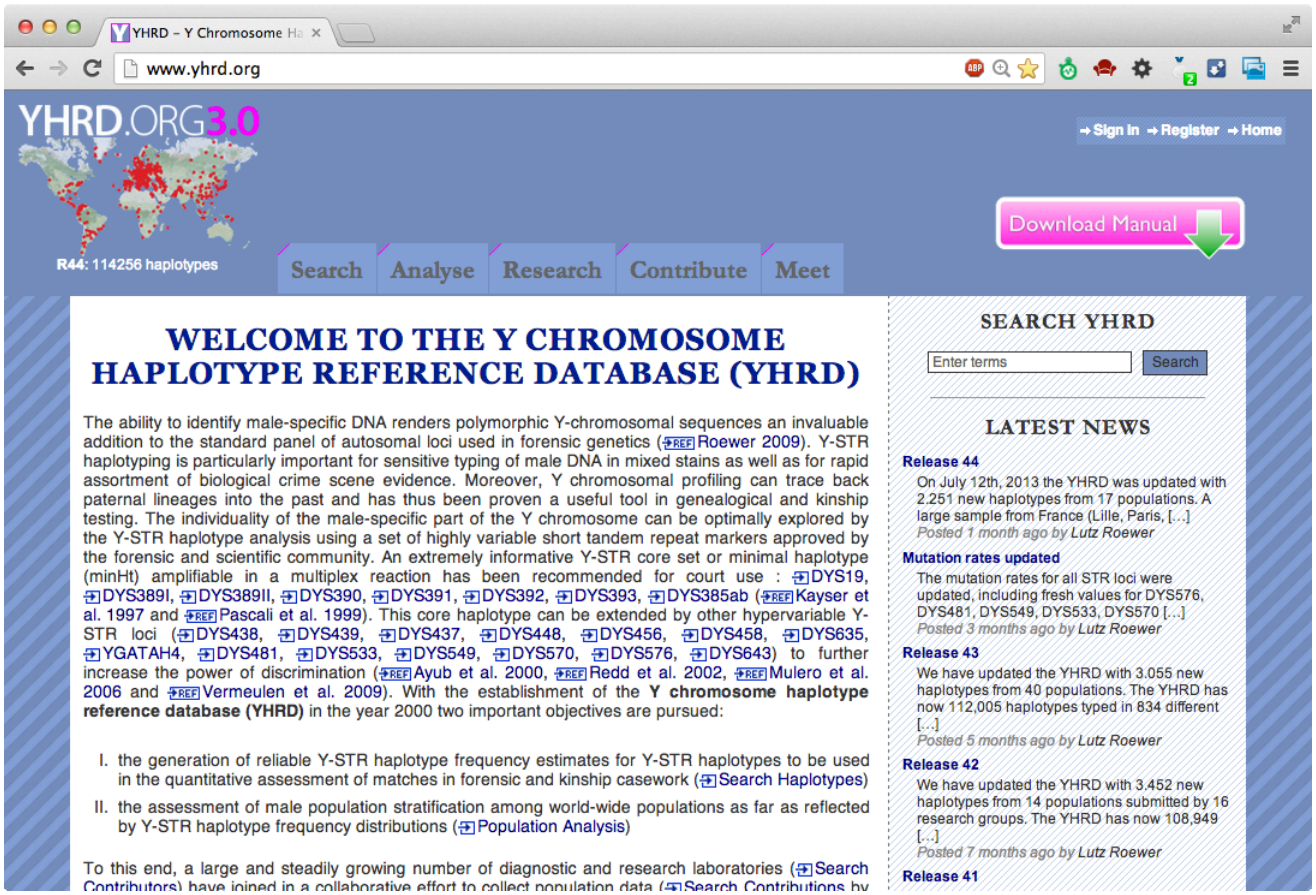


Figure 1: YHRD Homepage

2 Current state of the database

By September 2013 allmost 115,000 9-locus Haplotypes (including almost 56,000 YFiler Haplotypes) (see Figure 2) from 851 sampling locations in 113 countries have been submitted by 237 institutes and laboratories. In geographic terms, about 39% of the YHRD samples are from Europe, 32% from Asia, 17% from South America, 6% from North America, 4% from Africa and 2% from Oceania/Australia (see Figure 3).

We continuously receive new data from submitters (haplotypes, mutation rates etc.) and update the YHRD regularly. Since new releases replace the previous ones, the release number and date is an important part of the "Search result" document. The Release Notes are included in the bottom line of all documents.

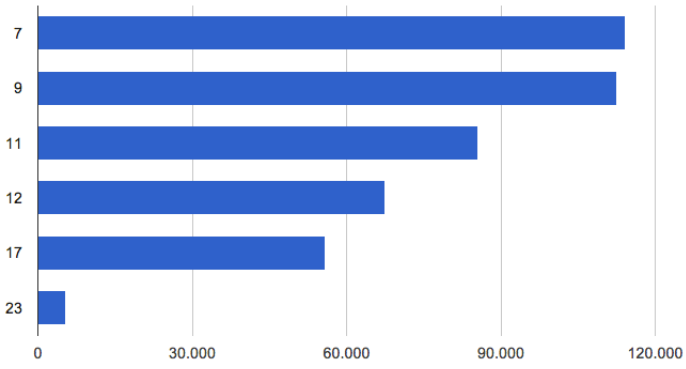


Figure 2: Distribution of different Haplotype formats

3 Navigation through the Website

3.1 Search the Database – `http://www.yhrd.org/Search`

3.1.1 By Haplotype – `http://www.yhrd.org/Search/Haplotype`

The database can be searched for all single alleles and all allele combinations (see Figure 4). The database supports the most frequently used Haplotype formats (e.g. Minimal, SWGDAM, Powerplex, YFiler and Y23 Haplotypes) for which differently-sized databases exist. At each position of the mask the respective allele must be entered (see Figure 5). The main sets of Markers are separated in different views (YHRD Standard see Figure 6, PowerPlex Y, Yfiler see Figure 7, PowerPlex Y23 see Figure 8) for easier input. To facilitate the search the scroll-down menu shows all alleles which have been observed in the database at the respective locus yet (see Figure 5).

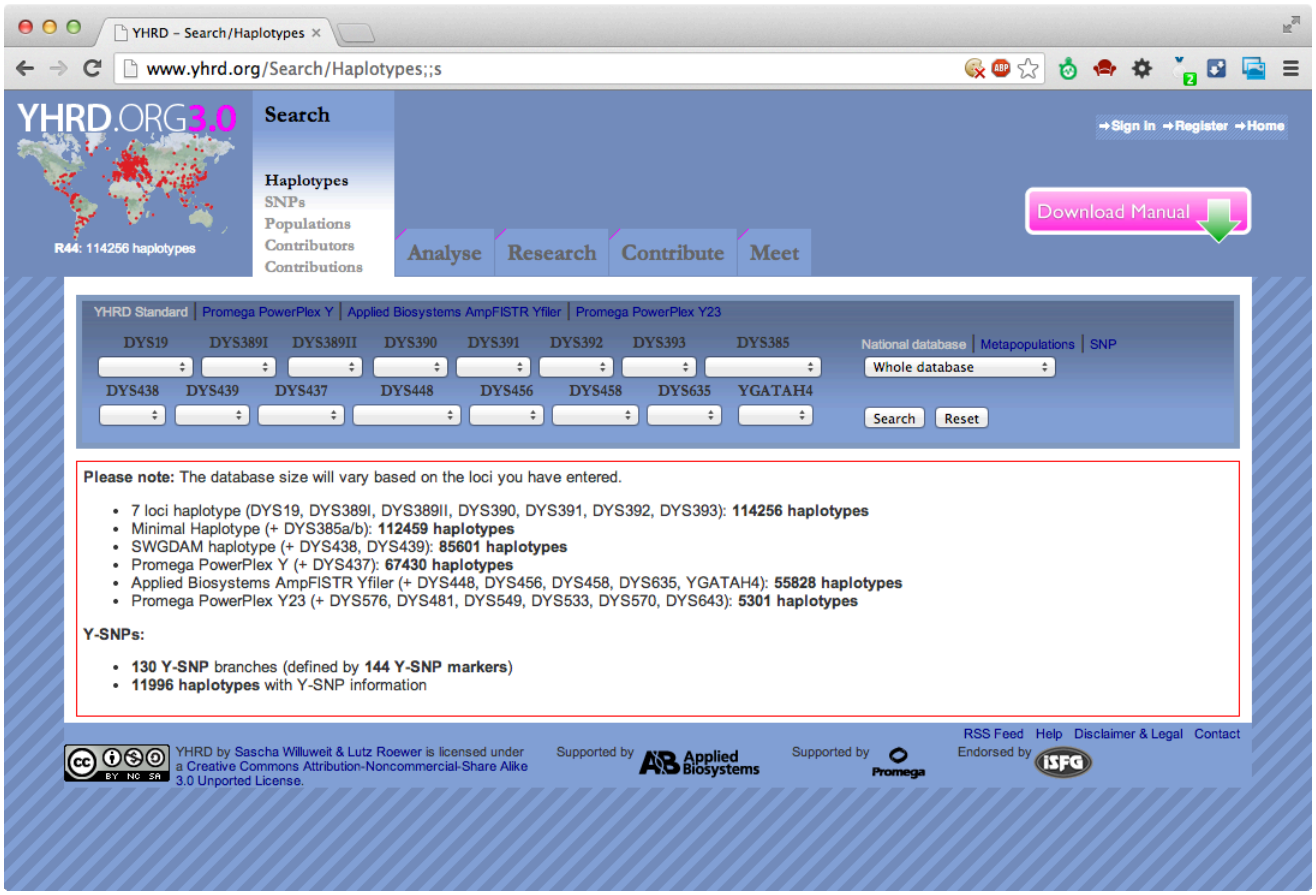


Figure 4: Search Haplotypes: Start

Pay attention to irregularly-spaced alleles which are named according to the ISFG recommendations (Gusmão et al. (2006)). Alleles which are irregular at a certain locus are elsewhere documented[6]. The upper line displays the 9 loci of the minimal Haplotype (see Figure 6). Nearly all Y chromosomes of the database are typed for the minimal Haplotype. The second line of loci includes all other loci up to maximal 17. About 28% of all Y chromosomes in the database are typed with the extension of 17 loci (Yfiler). Clicking "Reset" empties the search mask.

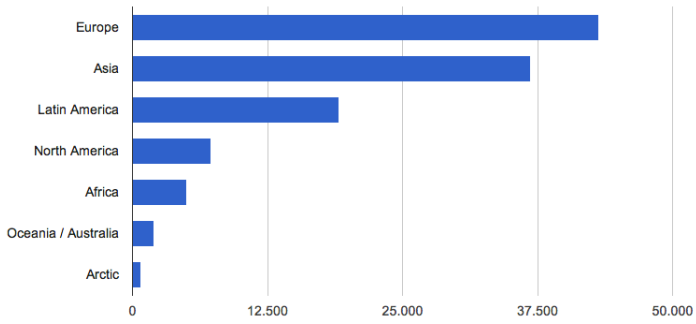


Figure 3: Continental distribution of Haplotypes

The database has a structure which can be adapted to your needs. Please choose in the mask the required database, either the whole database or a Metapopulation database (e.g. Western European database, see Figure 9), a national database (e.g. Argentina, see Figure 10) or Y-SNP defined subset of the database (see Figure 11)). Currently, you can choose from 34 meta-populations (structured according to scientifically evidenced hierarchy, see 2) or 113 national databases (from Afghanistan to Yemen). If you choose the default database (Whole database) the result table shows the average frequency of the Haplotype in all Metapopulations and the confidence intervals. In addition, the number of single populations with matches is shown. By clicking the arrow to the left subdivisions of Metapopulations with match-numbers become visible (see Figure 12).

The bar beneath the search mask allows to select a match statistic within Metapopulations, continents (see Figures 6, 12 and 14) or Haplogroups (see Figure 16). Of course, the previously chosen Metapopulation influences the presentation of matches at continents. In addition, a third button "Frequency surveying estimates" allows to extrapolate frequencies of rare or absent Haplotypes from a-priori distributions (Roewer et al. (2000), see Figure 15). This option is yet only feasible for full minimal Haplotypes and the Western, Eastern and Southeastern European Metapopulation. The Haplogroup tab is reflecting the Y-SNP distribution of the resulting haplotypes.

Scrolling further down shows a scalable and movable map on basis of the Google Maps software, with the rela-

[6] See `http://www.yhrd.org/Research/Loci`

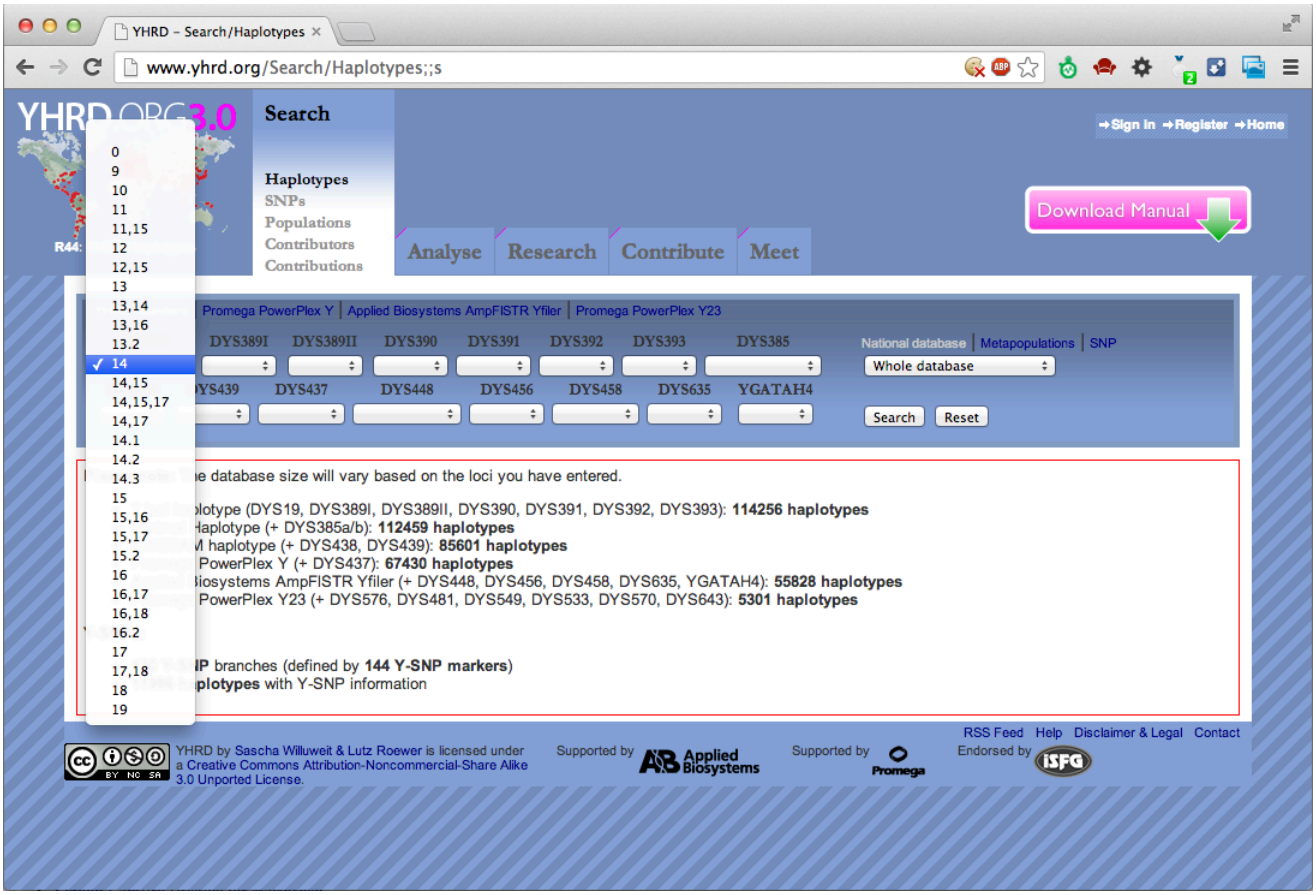


Figure 5: Search Haplotypes: Enter Haplotype

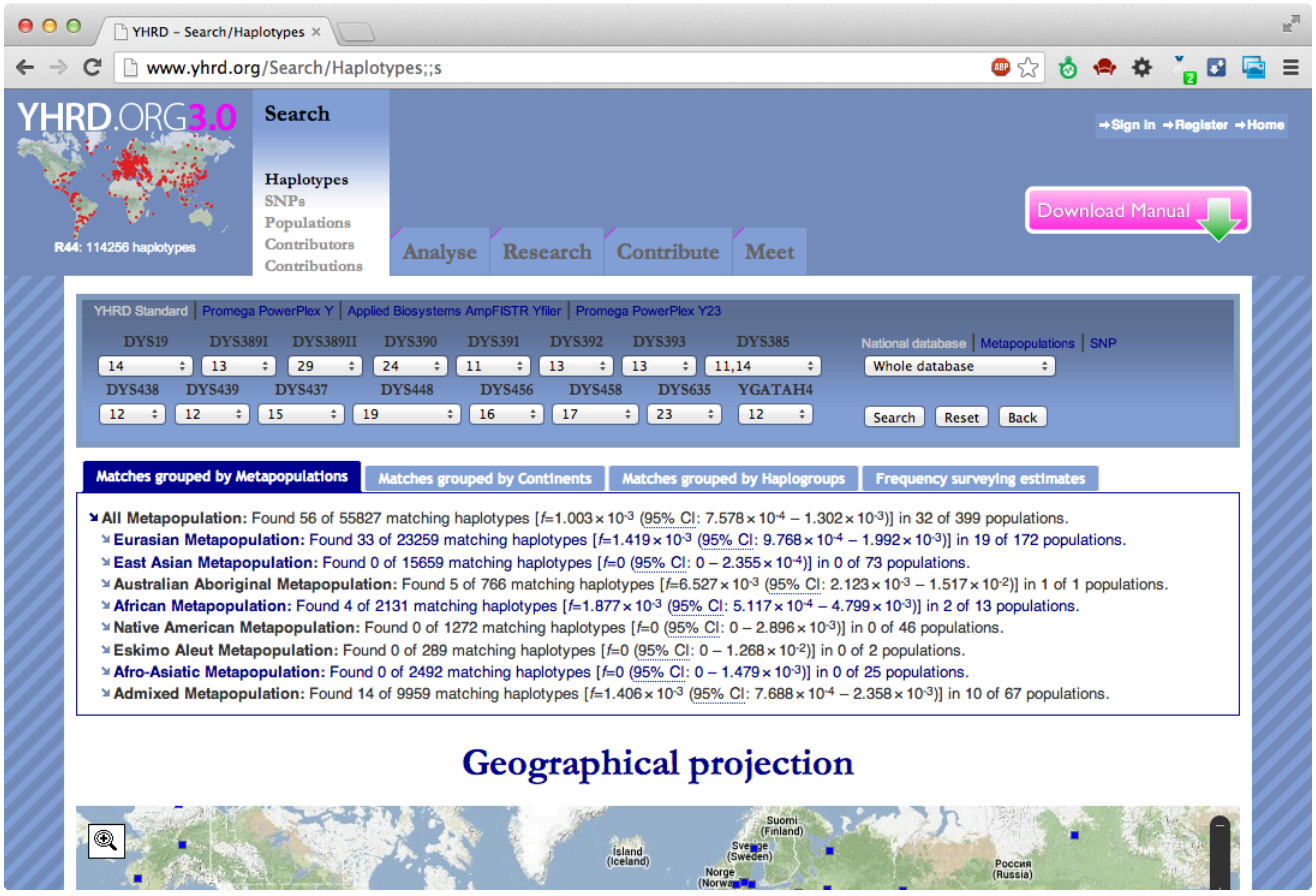


Figure 6: Search Haplotypes: YHRD Standard Haplotype input mask

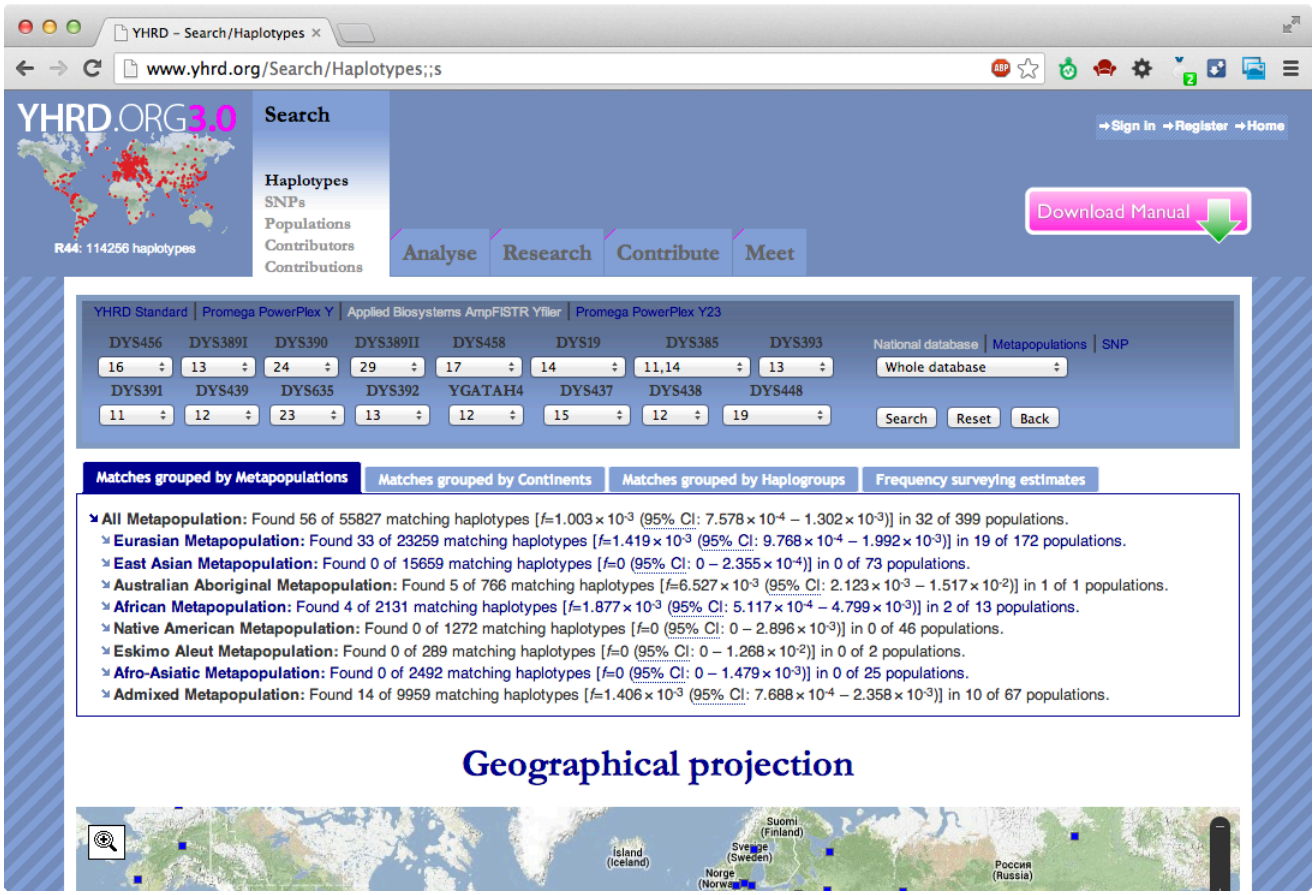


Figure 7: Search Haplotypes: Yfiler Haplotype input mask

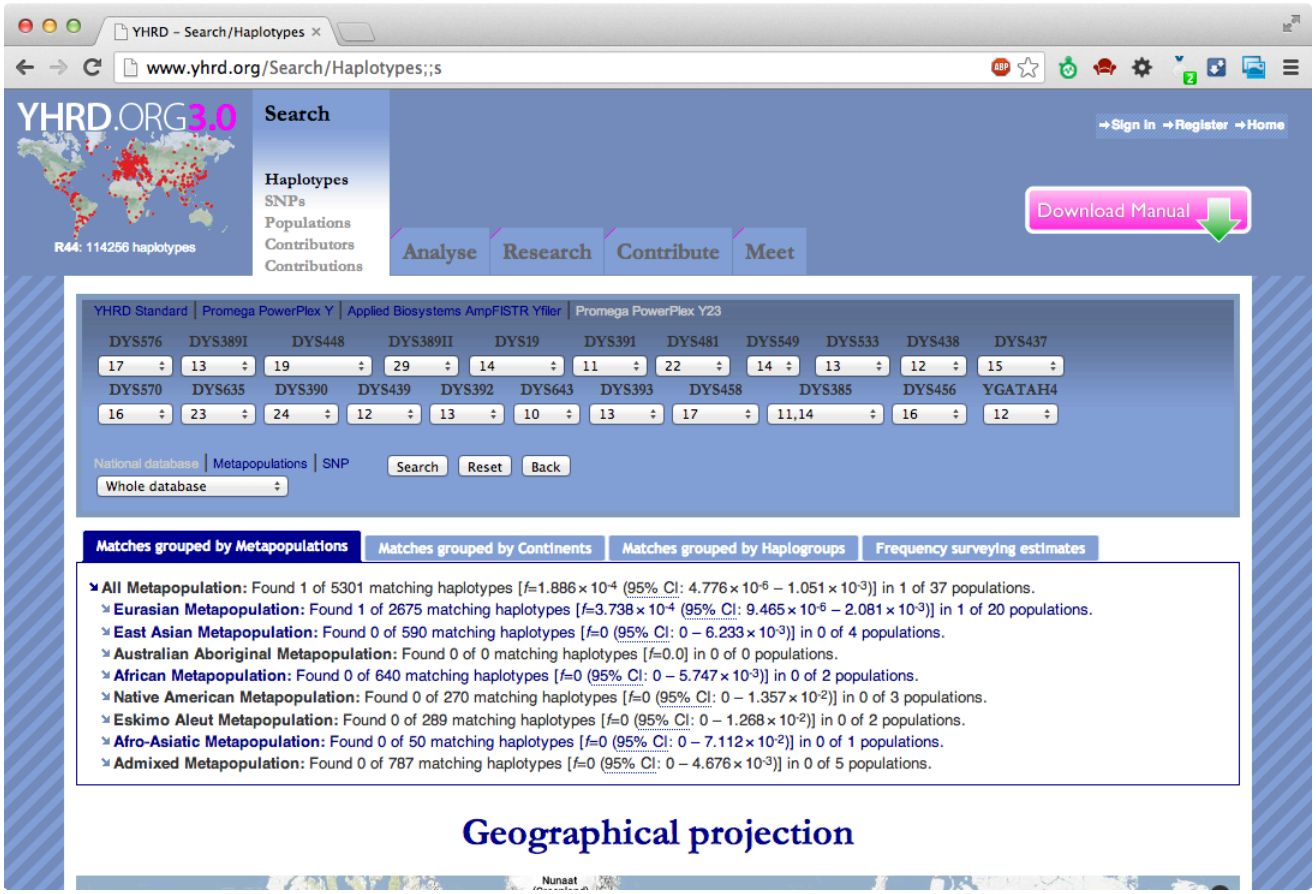


Figure 8: Search Haplotypes: PowerPlex Y23 Haplotype input mask

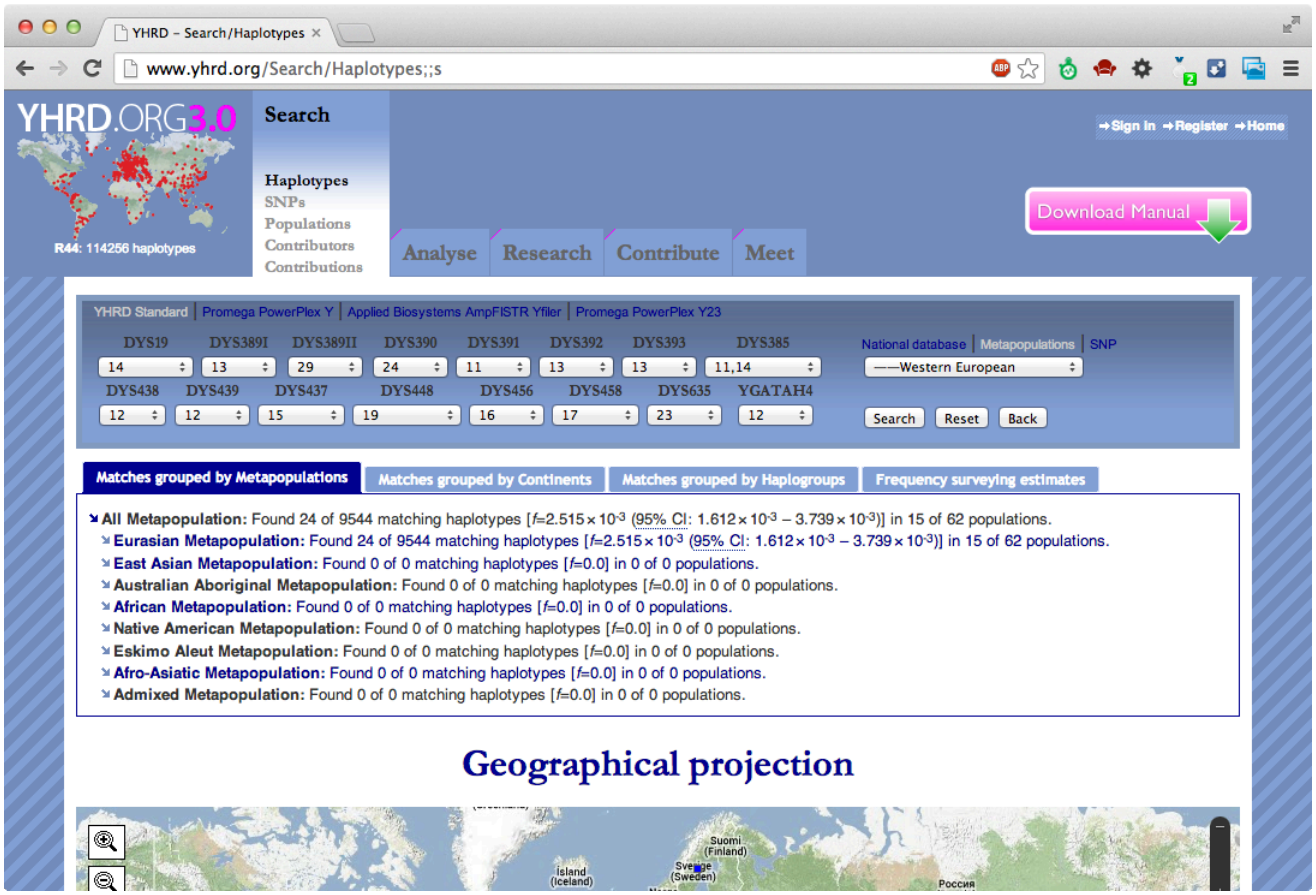


Figure 9: Search Haplotypes: Search in Metapopulations

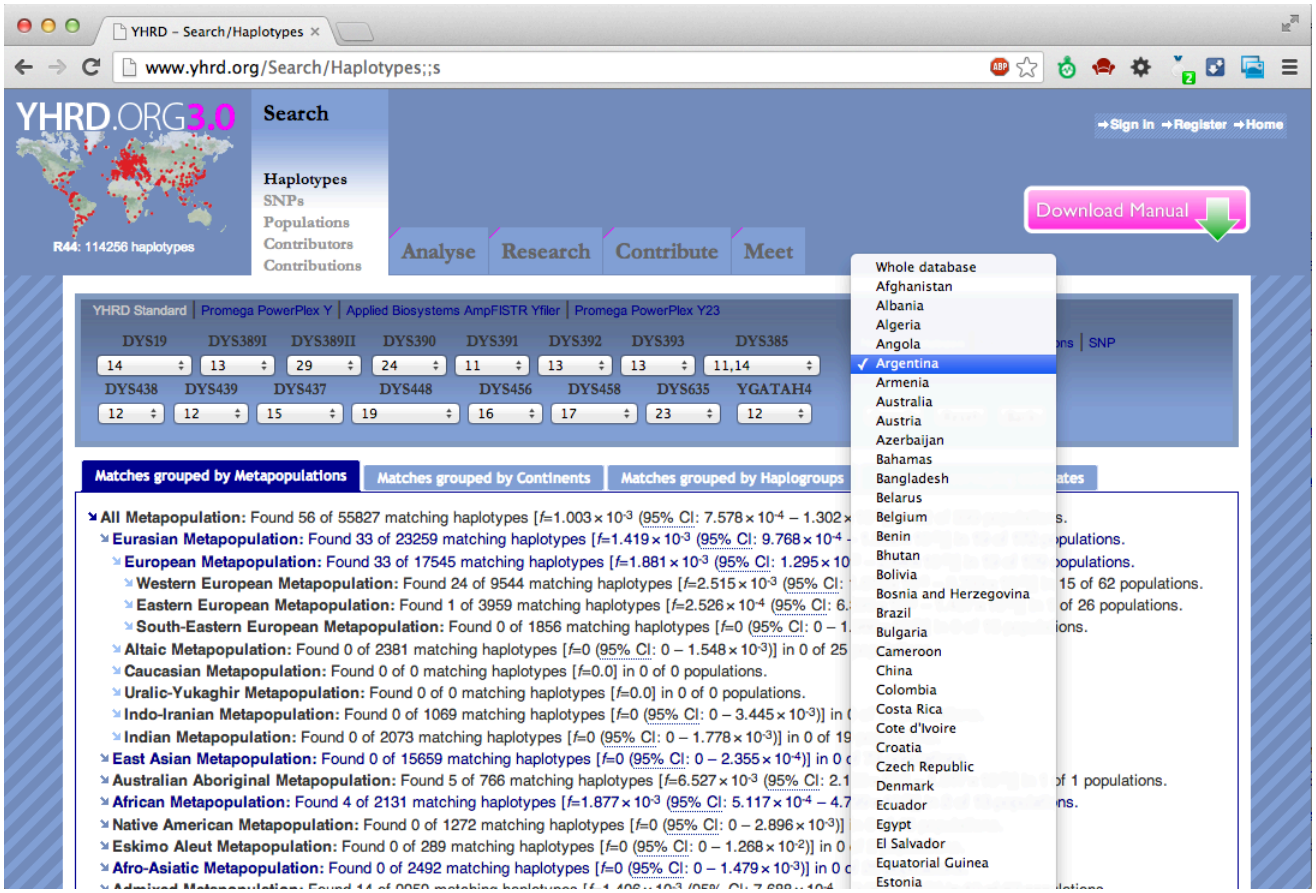


Figure 10: Search Haplotypes: Search in national databases

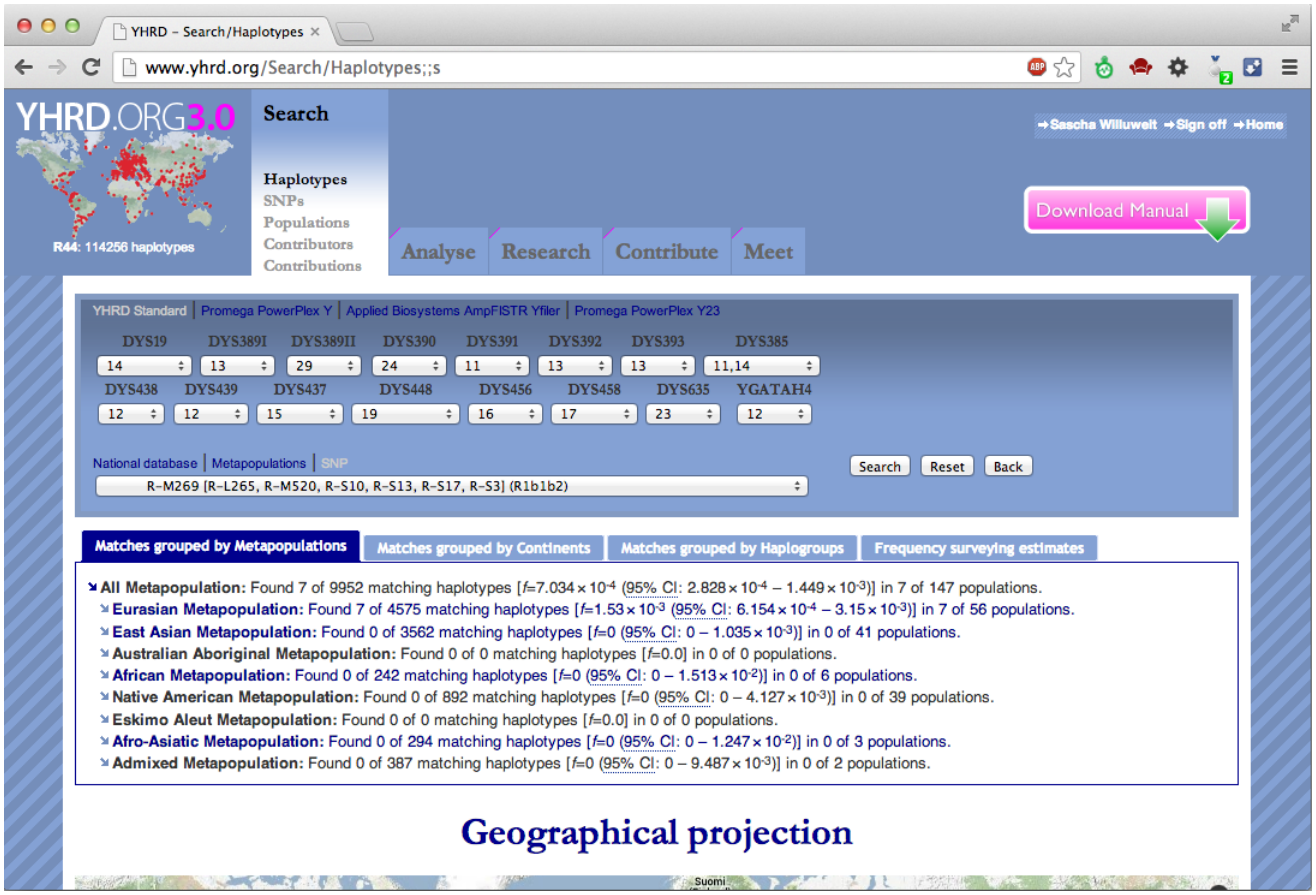


Figure 11: Search Haplotypes: Search in Y-SNP defined groups

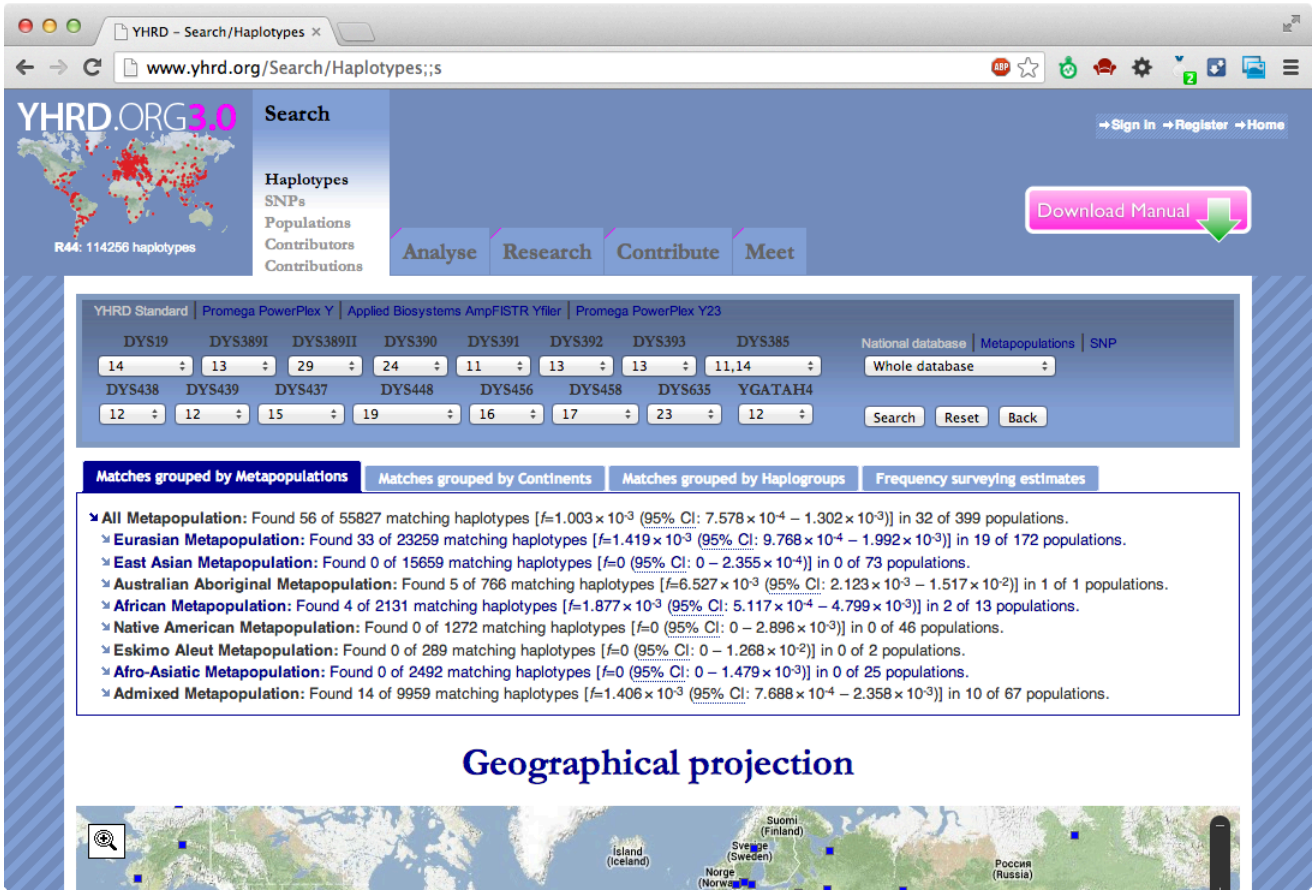


Figure 12: Search Haplotypes: Results by Metapopulation

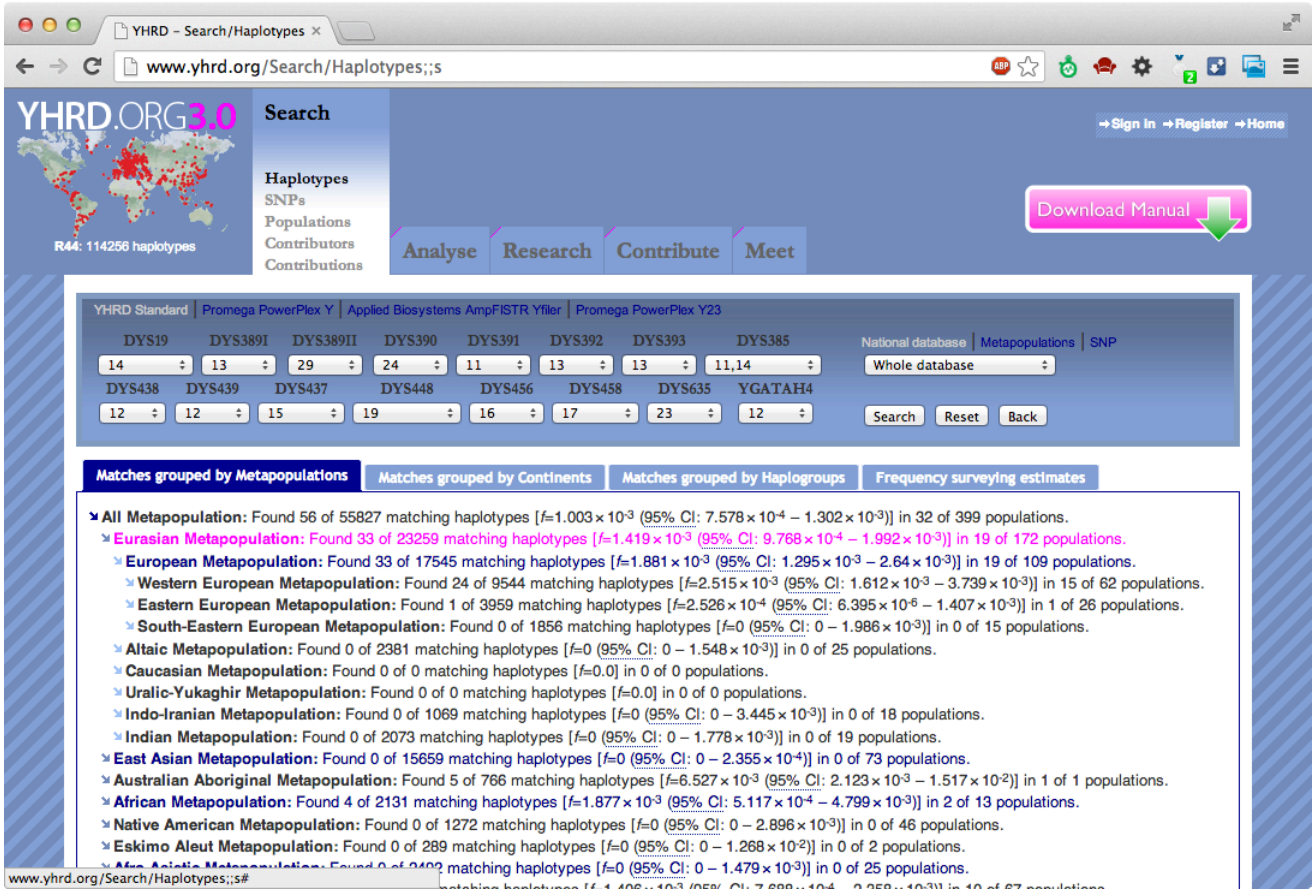


Figure 13: Search Haplotypes: Results by Metapopulation: Open hierarchy

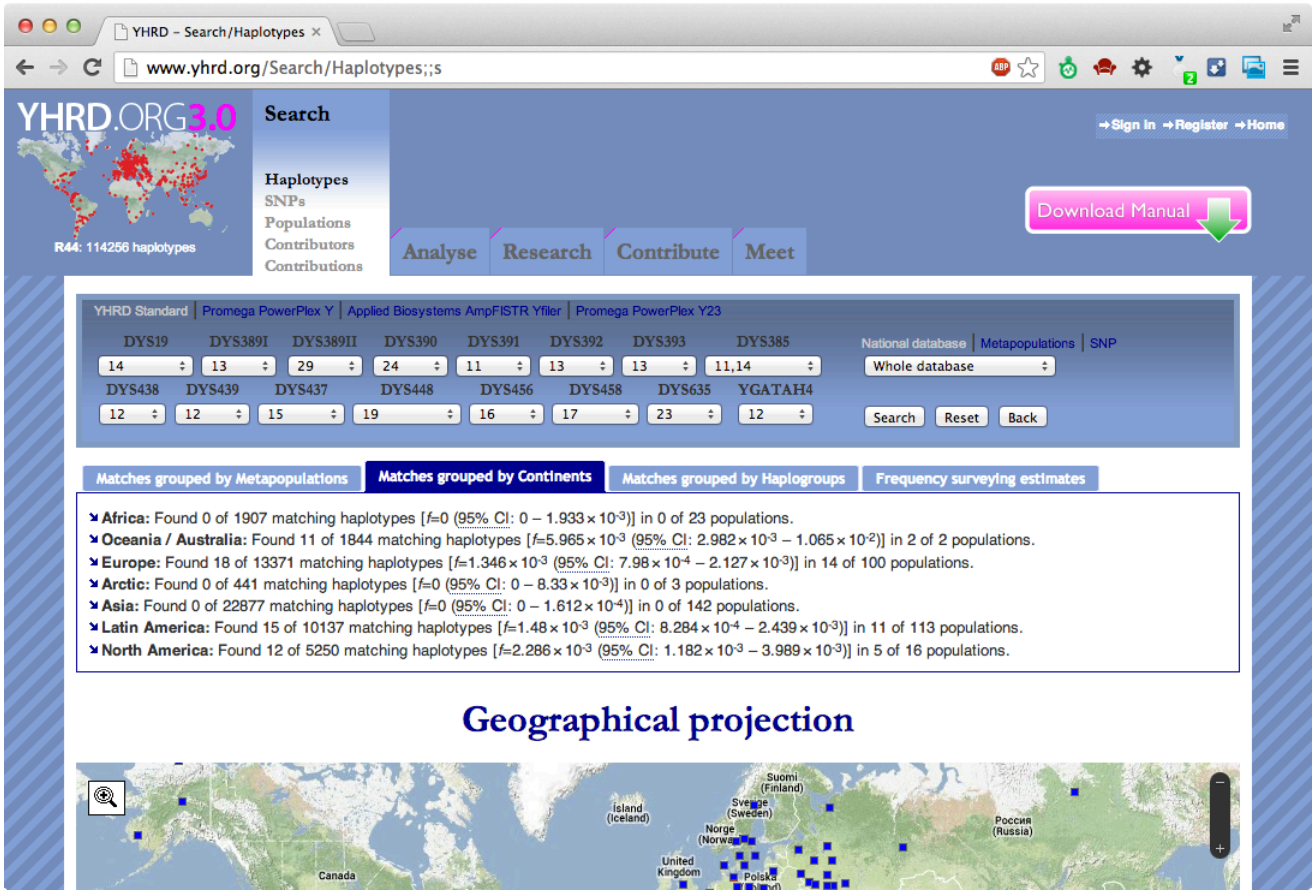


Figure 14: Search Haplotypes: Results by Continent

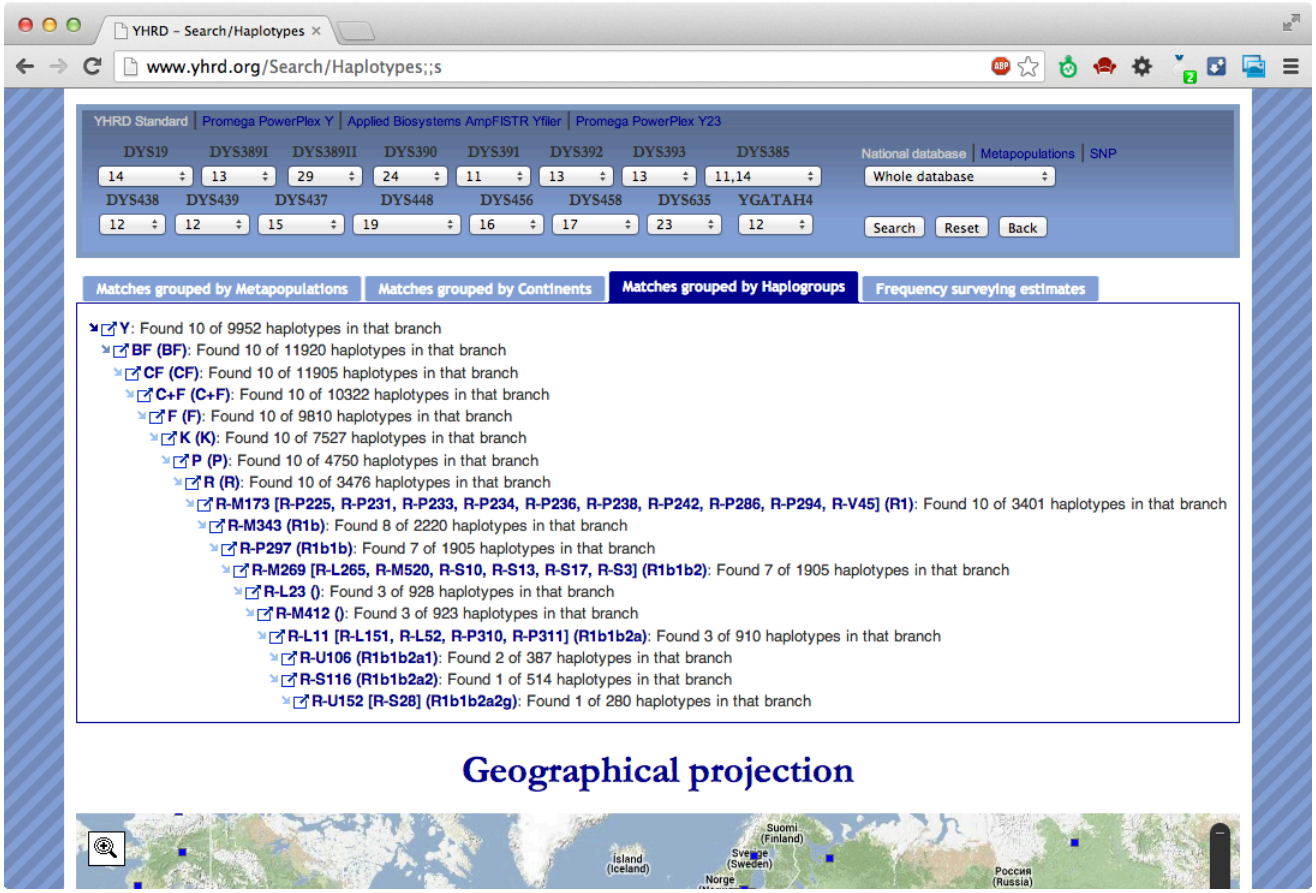


Figure 15: Search Haplotypes: Results by Haplogroup

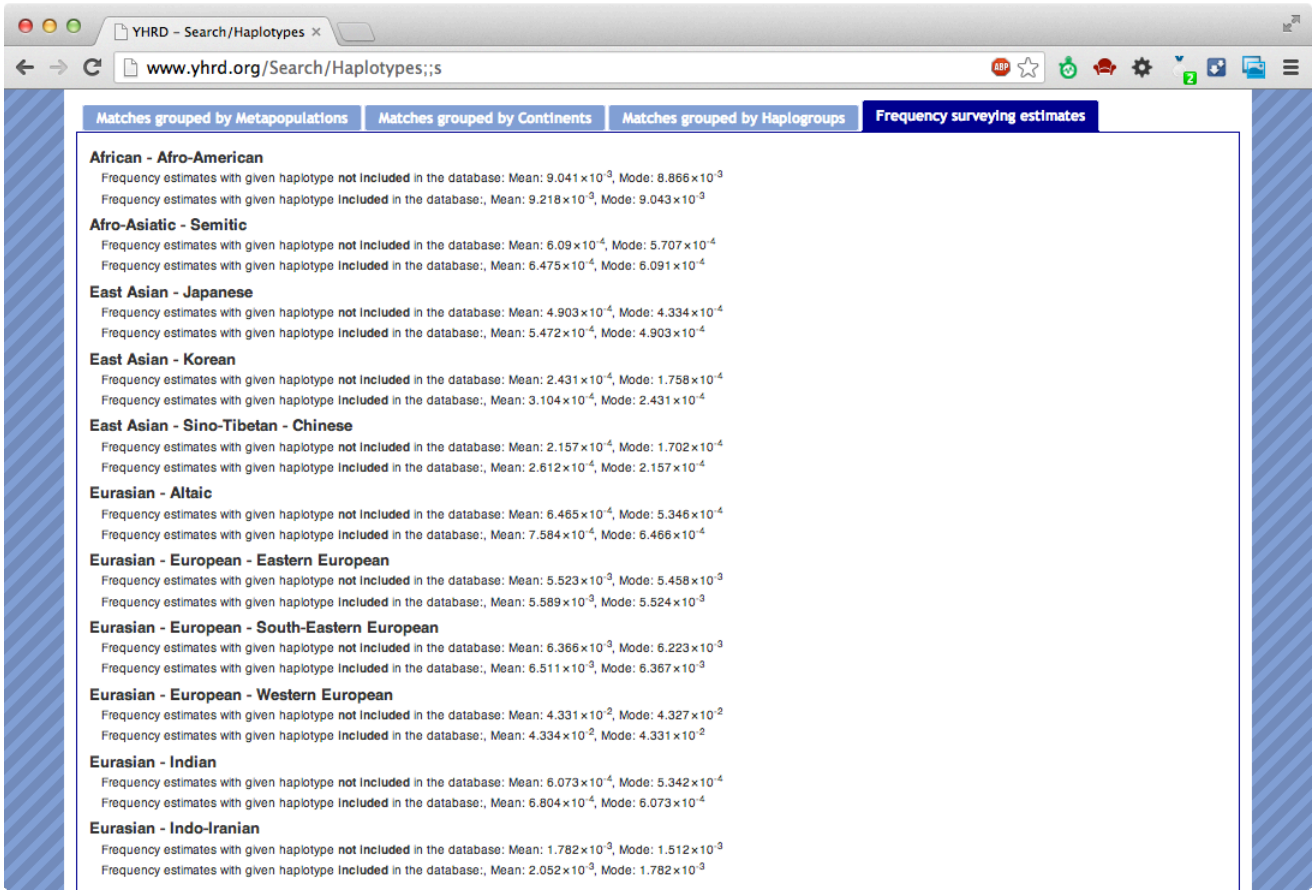


Figure 16: Search Haplotypes: Results by Frequeny Surveying

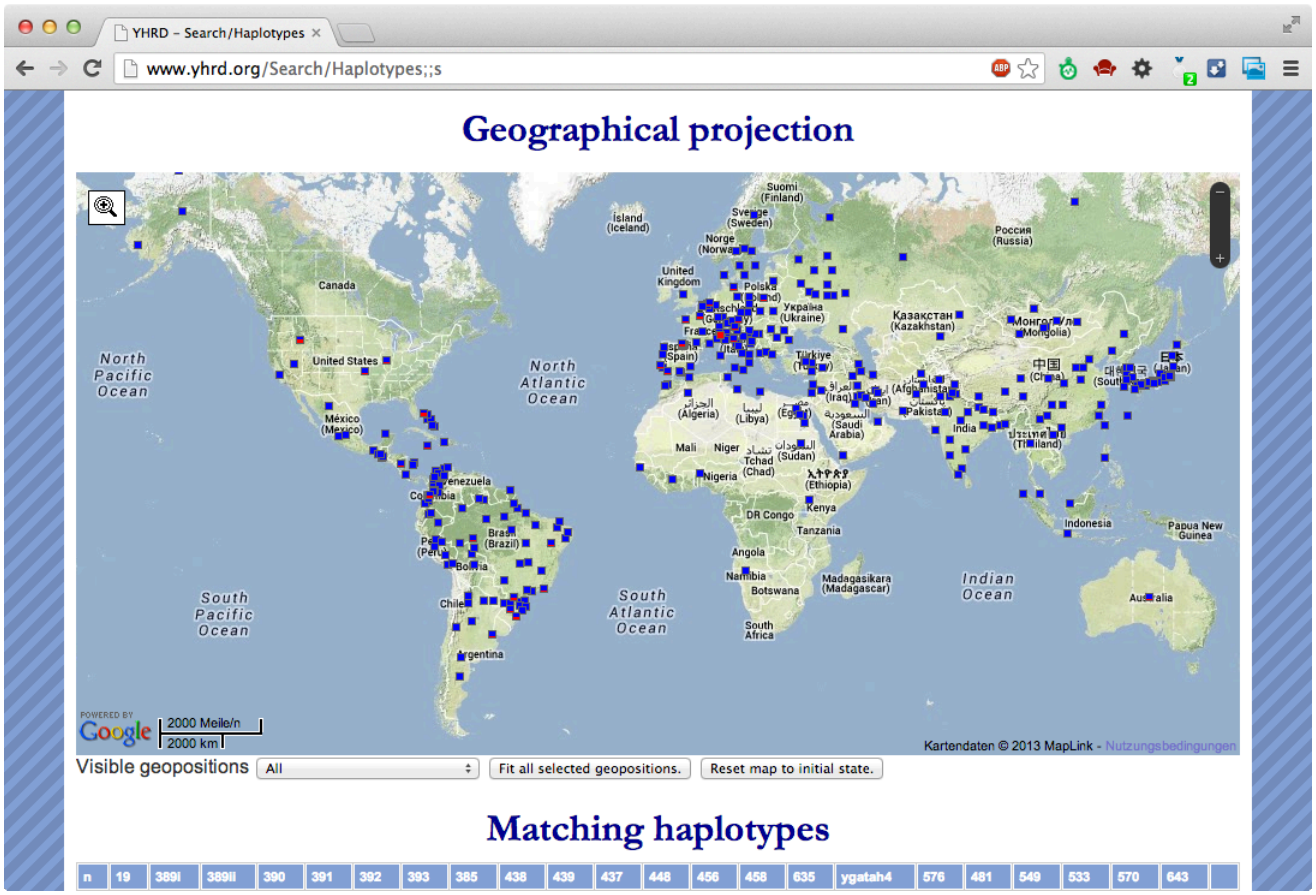


Figure 17: Search Haplotypes: Geographical projection

tive proportion of matches shown in red/blue for each Geoposition (see Figure 17). Its possible to select from a scroll-down menu all Geopositions either matched or non-matched, only those with matches or only those with a frequency between 20-80% of the maximal frequency (see Figure 19). There are additional tools for custom zooming the result map (see Figures 20, 21 and 22). Clicking a Geoposition in the map opens a window with essential information on the population sample, including name, assigned continent, ethnicity and Metapopulation as well as the frequency of the searched Haplotypes at this position. Please take in mind, that at a given Geoposition more than one population could have been sampled (see Figure 18).

Further down a table with all matched Haplotypes is shown ordered according to N (see Figure 23). Clicking to the headers changes the sorting order of the entries in the column (ascending or descending).
The next table shows all "Neighbour Haplotypes" and their frequencies with 1 repeat step up or down per locus (see Figure 24). This feature is very useful to find a cluster of evolutionary related Haplotypes surrounding the most frequent (or modal) Haplotypes of the cluster. By clicking the two little arrows on the right a new search is enabled for the respective neighboring Haplotype.

The population summary beneath these tables finally presents the matches per single population sample ordered (in default modus) by absolute numbers (see Figure 25). Again, clicking the headers changes the criterion of sorting order. Clicking in the population name opens a window with essential information on the contributor(s) or submitting institutions (see Figure 26). The accession number for this population can be clicked in this window and guides you directly to the "Contributors" page. Clicking the "Reference" opens the reference with further information.

At the bottom of the "Result page" you'll find a release note (see Figure 27) with the date and number of the current version of the database as well as the filling state.
Nomenclature for all loci follows the ISFG recommendations (Gusmão et al. (2006)). Please note, that the nomenclature for the GATA H4 locus as defined by the allelic ladder in the YFiler kit does not follow this recommendation; to avoid errors introduced by conversion of one nomenclature to another we have decided to use the YFiler nomenclature at this locus for the YHRD.

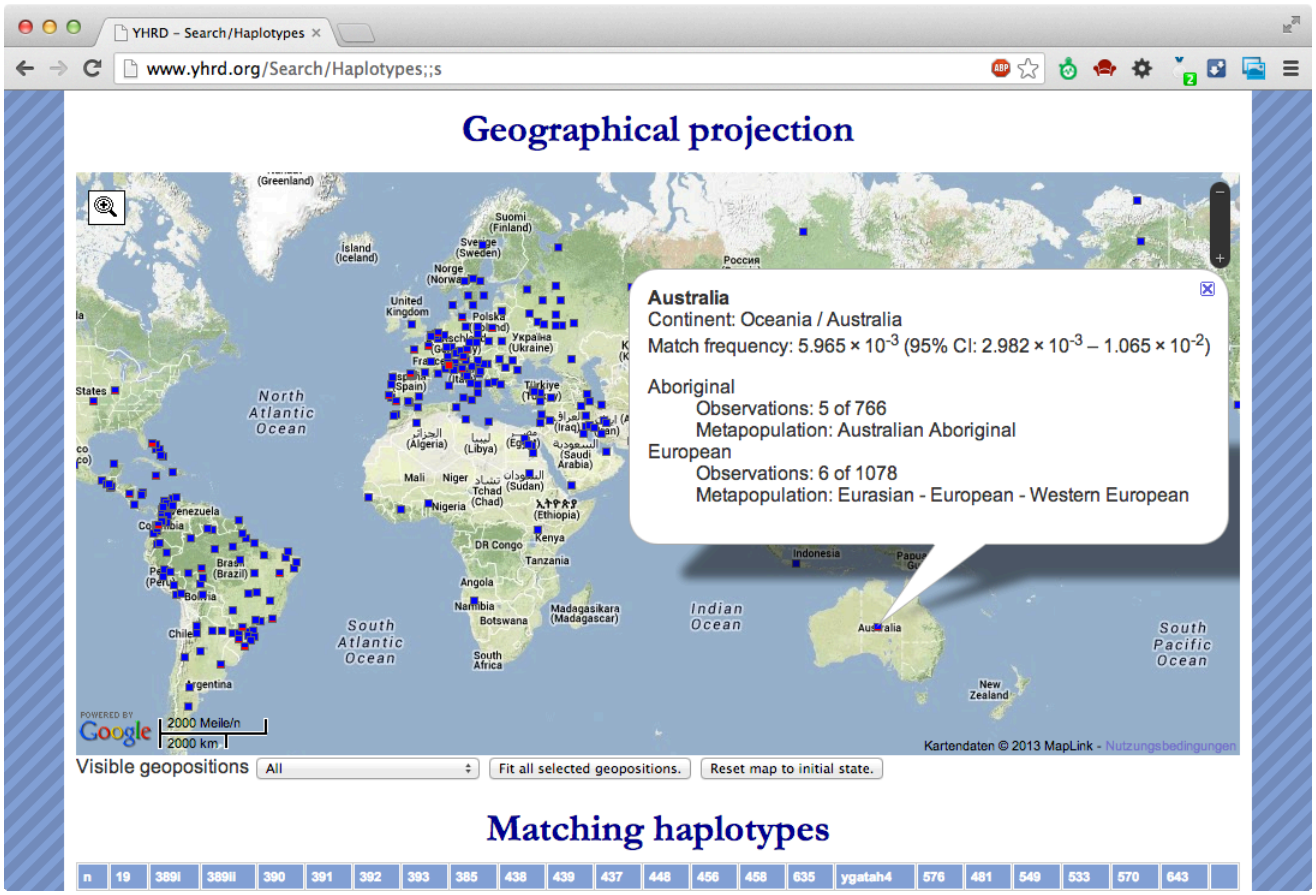


Figure 18: Search Haplotypes: Geographical projection: Geoposition information window

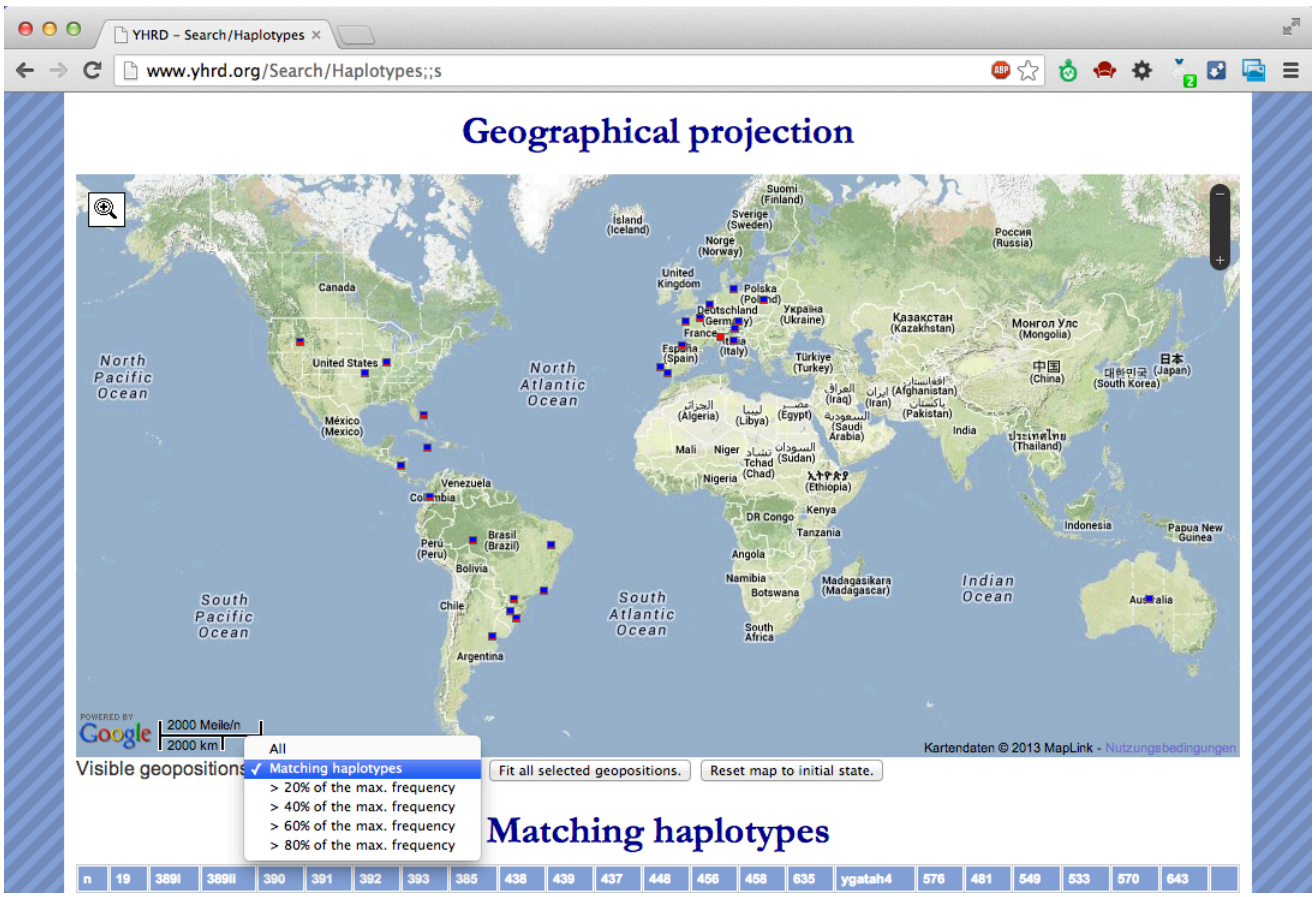


Figure 19: Search Haplotypes: Geographical projection: Show only populations with matches

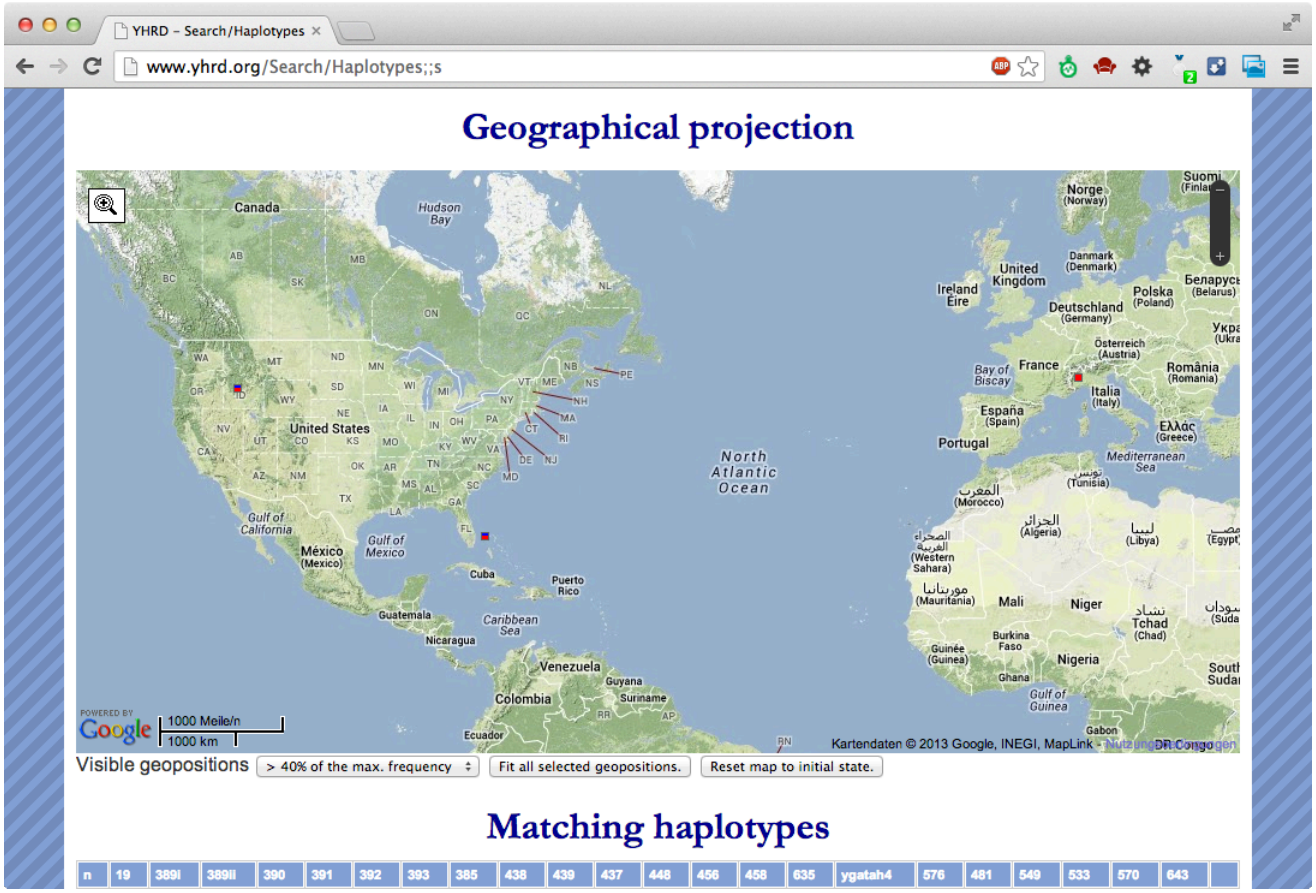


Figure 20: Search Haplotypes: Geographical projection: Auto-Zoom and Auto-Fit

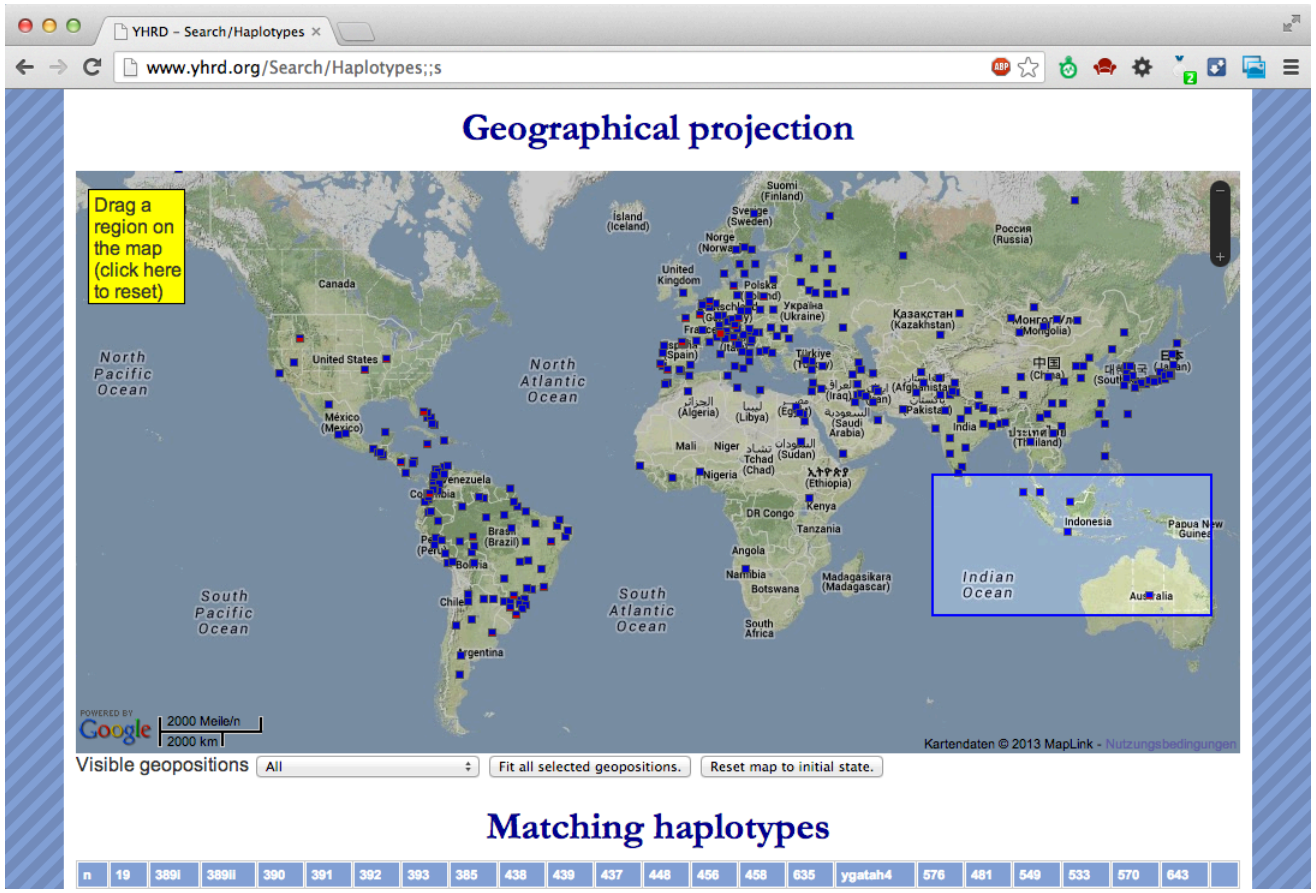


Figure 21: Search Haplotypes: Geographical projection: Zoom tool

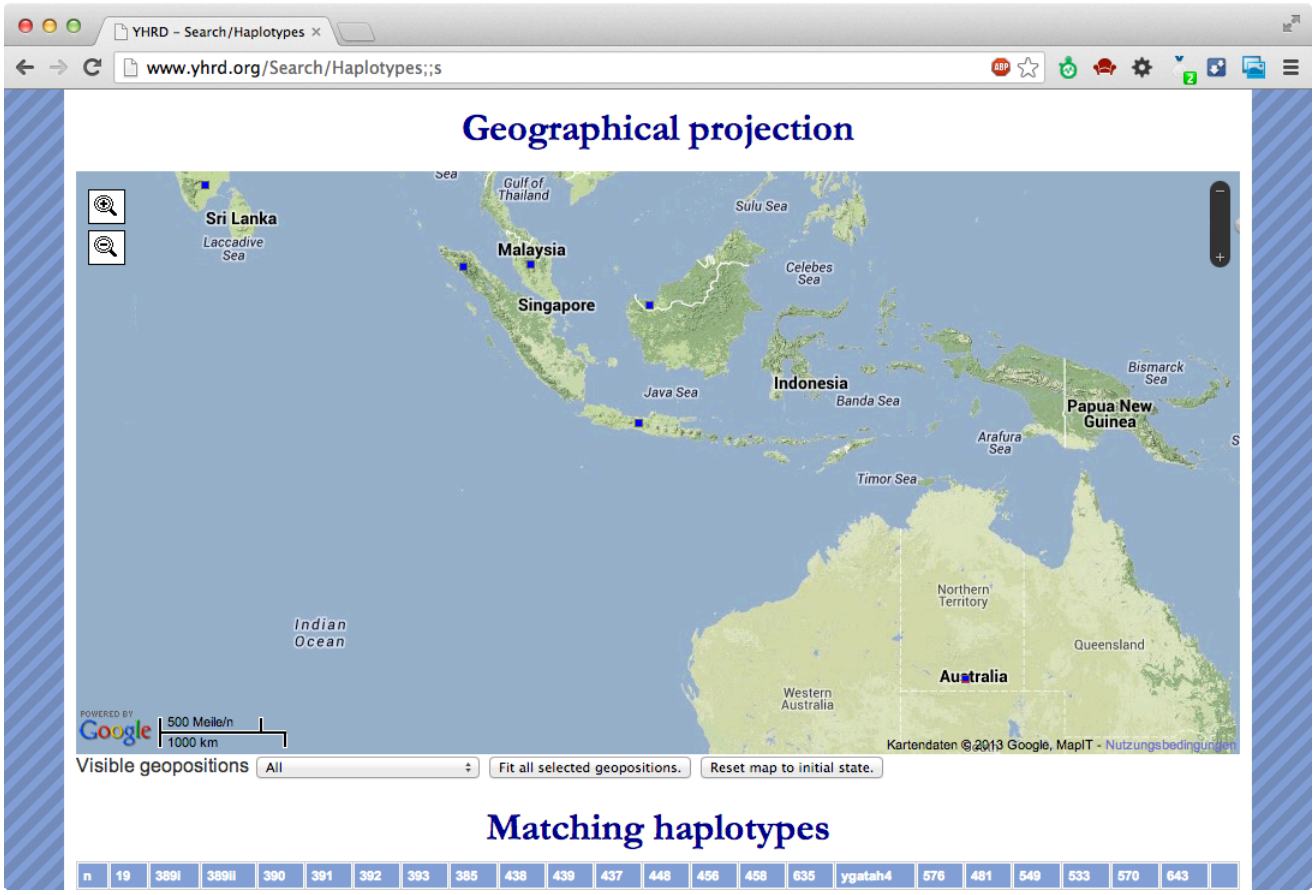


Figure 22: Search Haplotypes: Geographical projection: Zoom tool: Result

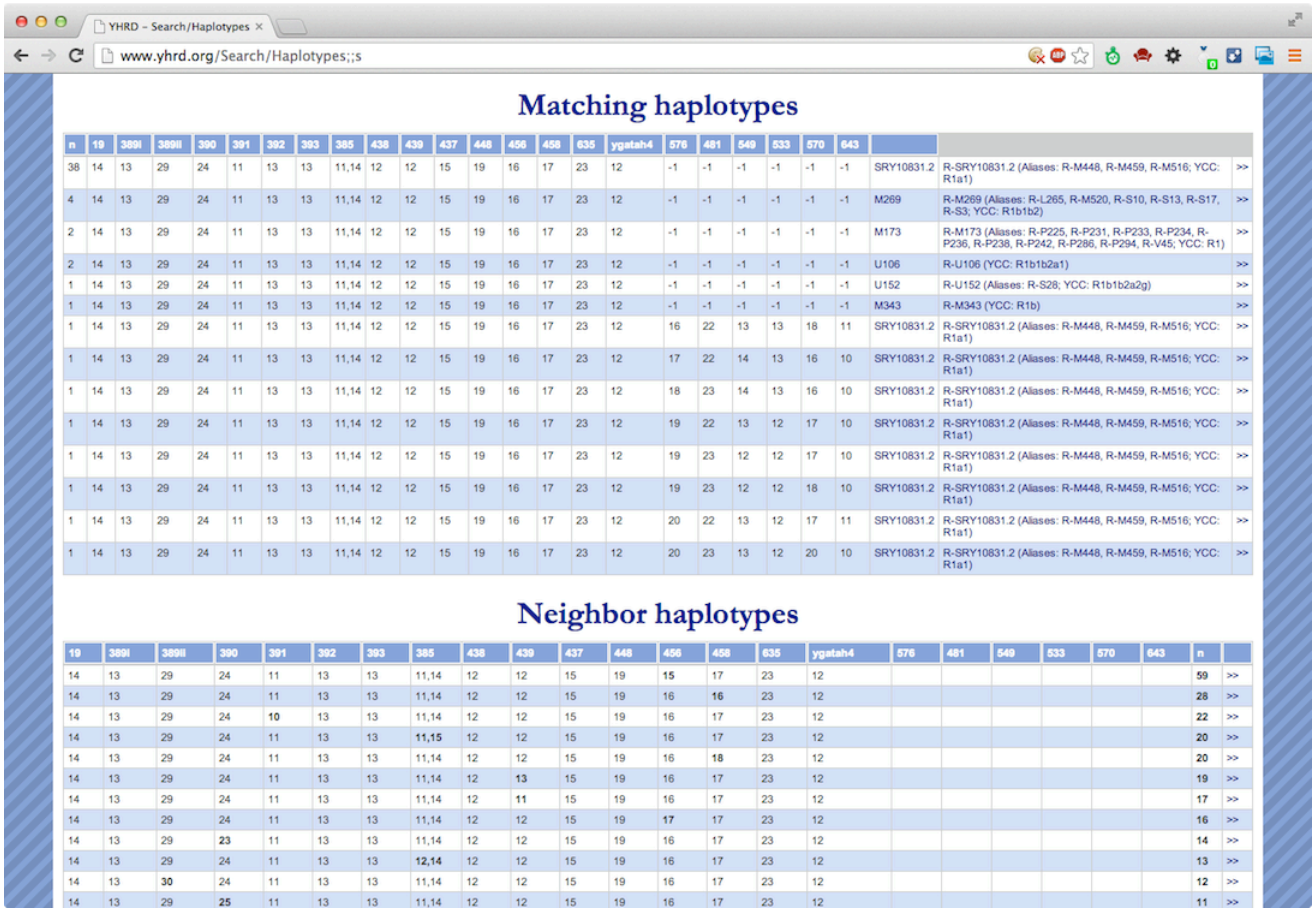
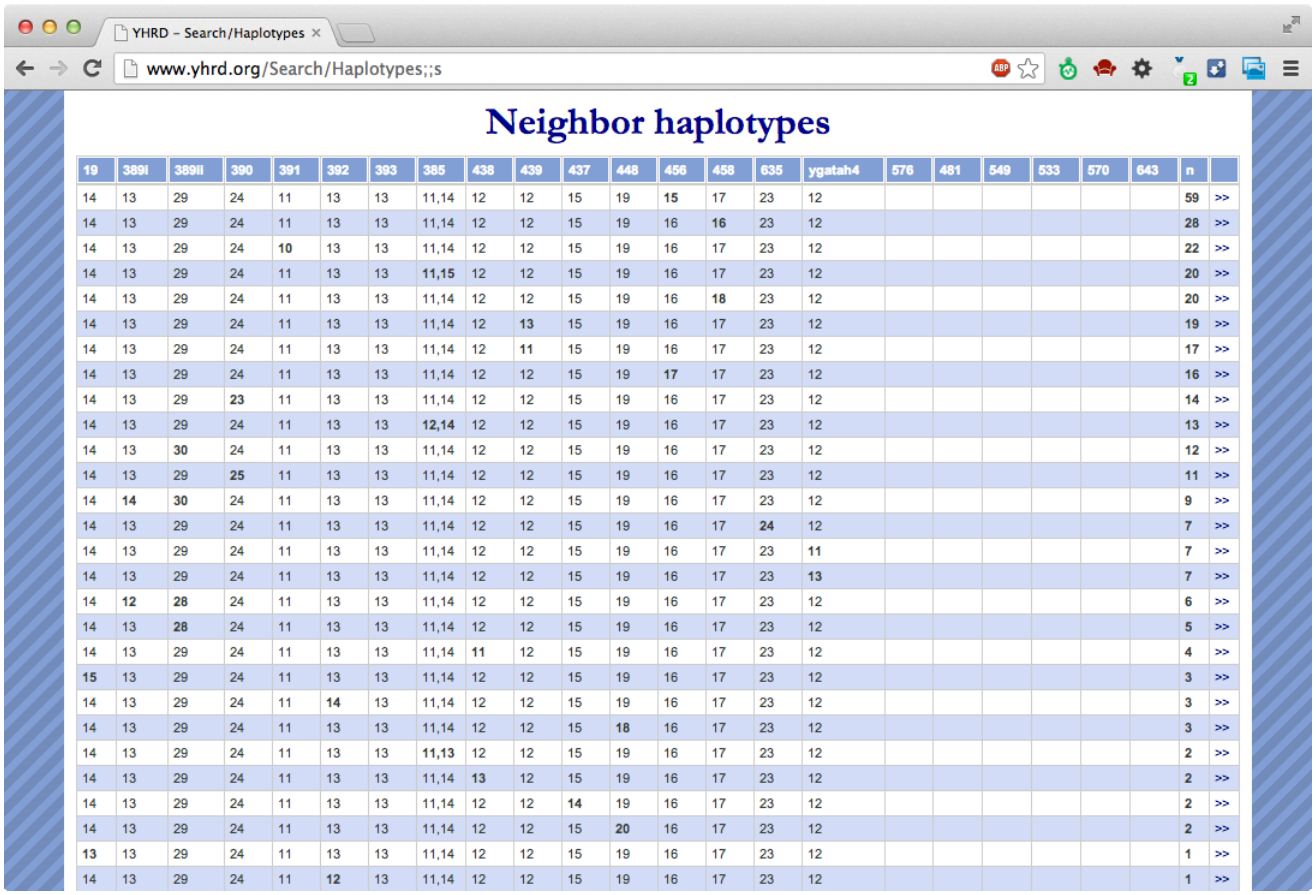


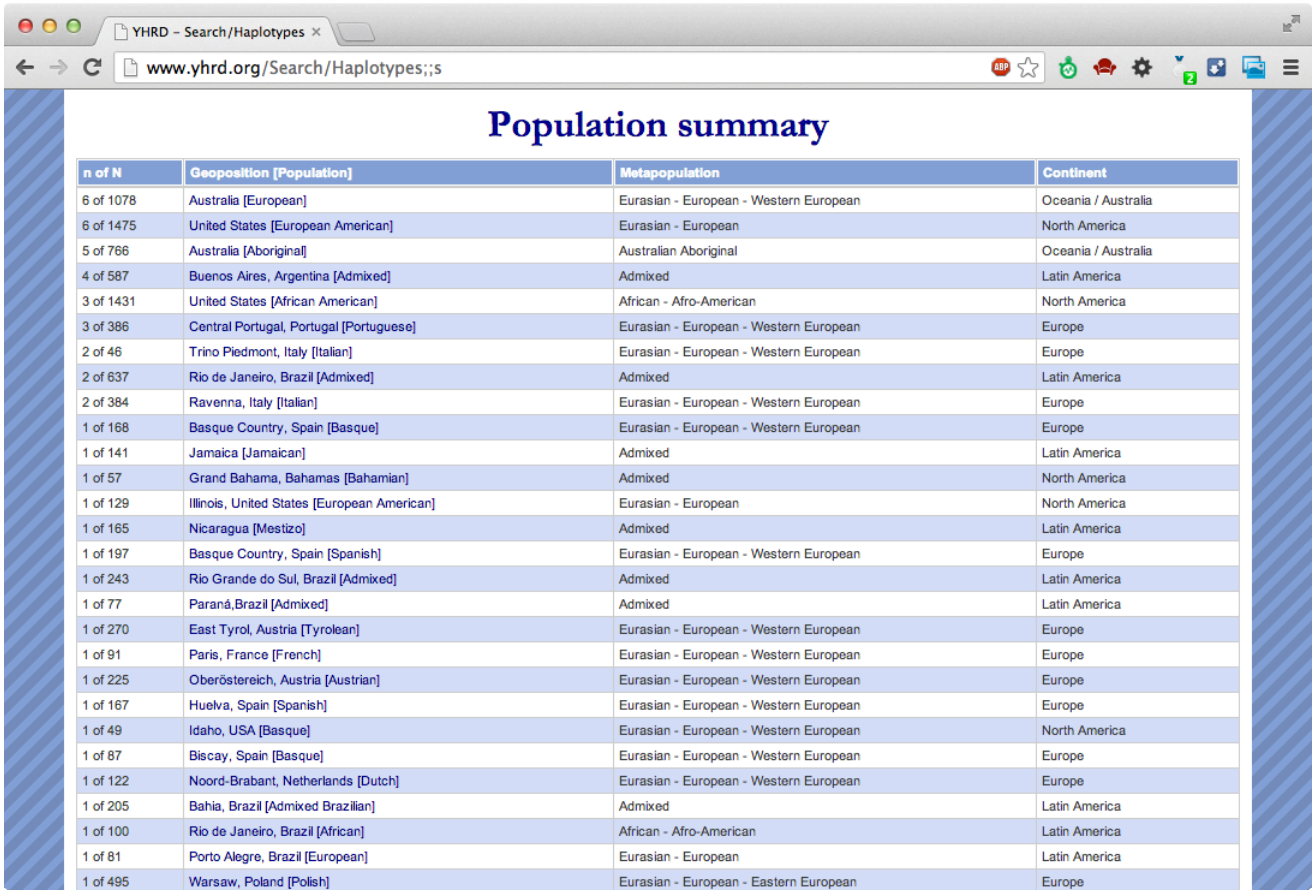
Figure 23: Search Haplotypes: Matching Haplotypes



Neighbor haplotypes

19	389I	389II	390	391	392	393	385	438	439	437	448	456	458	635	ygatah4	576	481	549	533	570	643	n	
14	13	29	24	11	13	13	11,14	12	12	15	19	**15**	17	23	12							**59**	>>
14	13	29	24	11	13	13	11,14	12	12	15	19	16	**16**	23	12							**28**	>>
14	13	29	24	**10**	13	13	11,14	12	12	15	19	16	17	23	12							**22**	>>
14	13	29	24	11	13	13	**11,15**	12	12	15	19	16	17	23	12							**20**	>>
14	13	29	24	11	13	13	11,14	12	12	15	19	16	**18**	23	12							**20**	>>
14	13	29	24	11	13	13	11,14	12	**13**	15	19	16	17	23	12							**19**	>>
14	13	29	24	11	13	13	11,14	12	**11**	15	19	16	17	23	12							**17**	>>
14	13	29	24	11	13	13	11,14	12	12	15	19	**17**	17	23	12							**16**	>>
14	13	29	**23**	11	13	13	11,14	12	12	15	19	16	17	23	12							**14**	>>
14	13	29	24	11	13	13	**12,14**	12	12	15	19	16	17	23	12							**13**	>>
14	13	**30**	24	11	13	13	11,14	12	12	15	19	16	17	23	12							**12**	>>
14	13	29	**25**	11	13	13	11,14	12	12	15	19	16	17	23	12							**11**	>>
14	**14**	**30**	24	11	13	13	11,14	12	12	15	19	16	17	23	12							**9**	>>
14	13	29	24	11	13	13	11,14	12	12	15	19	16	17	**24**	12							**7**	>>
14	13	29	24	11	13	13	11,14	12	12	15	19	16	17	23	**11**							**7**	>>
14	13	29	24	11	13	13	11,14	12	12	15	19	16	17	23	**13**							**7**	>>
14	**12**	**28**	24	11	13	13	11,14	12	12	15	19	16	17	23	12							**6**	>>
14	13	**28**	24	11	13	13	11,14	12	12	15	19	16	17	23	12							**5**	>>
14	13	29	24	11	13	13	11,14	**11**	12	15	19	16	17	23	12							**4**	>>
15	13	29	24	11	13	13	11,14	12	12	15	19	16	17	23	12							**3**	>>
14	13	29	24	11	**14**	13	11,14	12	12	15	19	16	17	23	12							**3**	>>
14	13	29	24	11	13	13	11,14	12	12	15	**18**	16	17	23	12							**3**	>>
14	13	29	24	11	13	13	**11,13**	12	12	15	19	16	17	23	12							**2**	>>
14	13	29	24	11	13	13	11,14	**13**	12	15	19	16	17	23	12							**2**	>>
14	13	29	24	11	13	13	11,14	12	12	**14**	19	16	17	23	12							**2**	>>
14	13	29	24	11	13	13	11,14	12	12	15	**20**	16	17	23	12							**2**	>>
13	13	29	24	11	13	13	11,14	12	12	15	19	16	17	23	12							**1**	>>
14	13	29	24	11	**12**	13	11,14	12	12	15	19	16	17	23	12							**1**	>>

Figure 24: Search Haplotypes: Neighbour Haplotypes



Population summary

n of N	Geoposition [Population]	Metapopulation	Continent
6 of 1078	Australia [European]	Eurasian - European - Western European	Oceania / Australia
6 of 1475	United States [European American]	Eurasian - European	North America
5 of 766	Australia [Aboriginal]	Australian Aboriginal	Oceania / Australia
4 of 587	Buenos Aires, Argentina [Admixed]	Admixed	Latin America
3 of 1431	United States [African American]	African - Afro-American	North America
3 of 386	Central Portugal, Portugal [Portuguese]	Eurasian - European - Western European	Europe
2 of 46	Trino Piedmont, Italy [Italian]	Eurasian - European - Western European	Europe
2 of 637	Rio de Janeiro, Brazil [Admixed]	Admixed	Latin America
2 of 384	Ravenna, Italy [Italian]	Eurasian - European - Western European	Europe
1 of 168	Basque Country, Spain [Basque]	Eurasian - European - Western European	Europe
1 of 141	Jamaica [Jamaican]	Admixed	Latin America
1 of 57	Grand Bahama, Bahamas [Bahamian]	Admixed	North America
1 of 129	Illinois, United States [European American]	Eurasian - European	North America
1 of 165	Nicaragua [Mestizo]	Admixed	Latin America
1 of 197	Basque Country, Spain [Spanish]	Eurasian - European - Western European	Europe
1 of 243	Rio Grande do Sul, Brazil [Admixed]	Admixed	Latin America
1 of 77	Paraná,Brazil [Admixed]	Admixed	Latin America
1 of 270	East Tyrol, Austria [Tyrolean]	Eurasian - European - Western European	Europe
1 of 91	Paris, France [French]	Eurasian - European - Western European	Europe
1 of 225	Oberösterreich, Austria [Austrian]	Eurasian - European - Western European	Europe
1 of 167	Huelva, Spain [Spanish]	Eurasian - European - Western European	Europe
1 of 49	Idaho, USA [Basque]	Eurasian - European - Western European	North America
1 of 87	Biscay, Spain [Basque]	Eurasian - European - Western European	Europe
1 of 122	Noord-Brabant, Netherlands [Dutch]	Eurasian - European - Western European	Europe
1 of 205	Bahia, Brazil [Admixed Brazilian]	Admixed	Latin America
1 of 100	Rio de Janeiro, Brazil [African]	African - Afro-American	Latin America
1 of 81	Porto Alegre, Brazil [European]	Eurasian - European	Latin America
1 of 495	Warsaw, Poland [Polish]	Eurasian - European - Eastern European	Europe

Figure 25: Search Haplotypes: Population summary

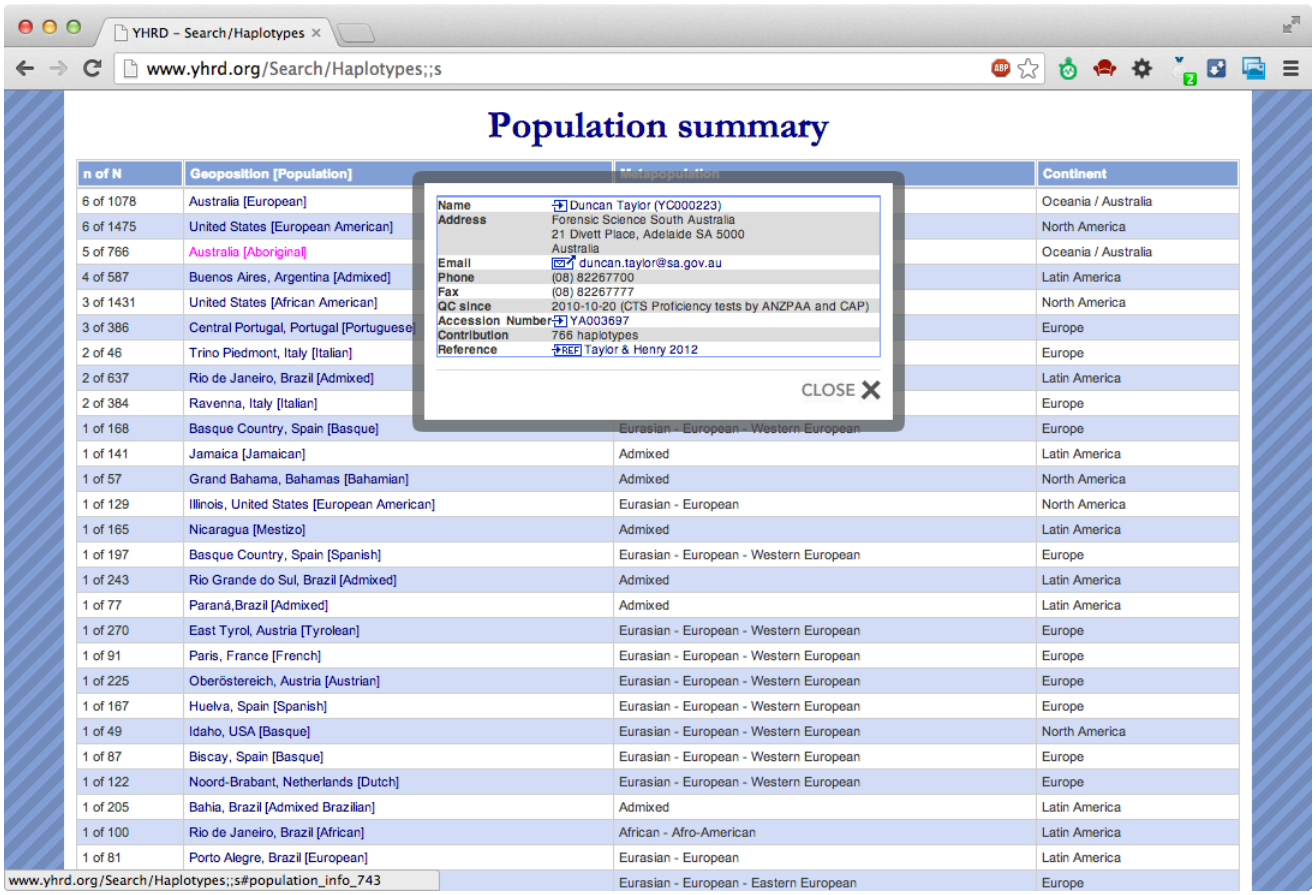


Figure 26: Search Haplotypes: Contributor information

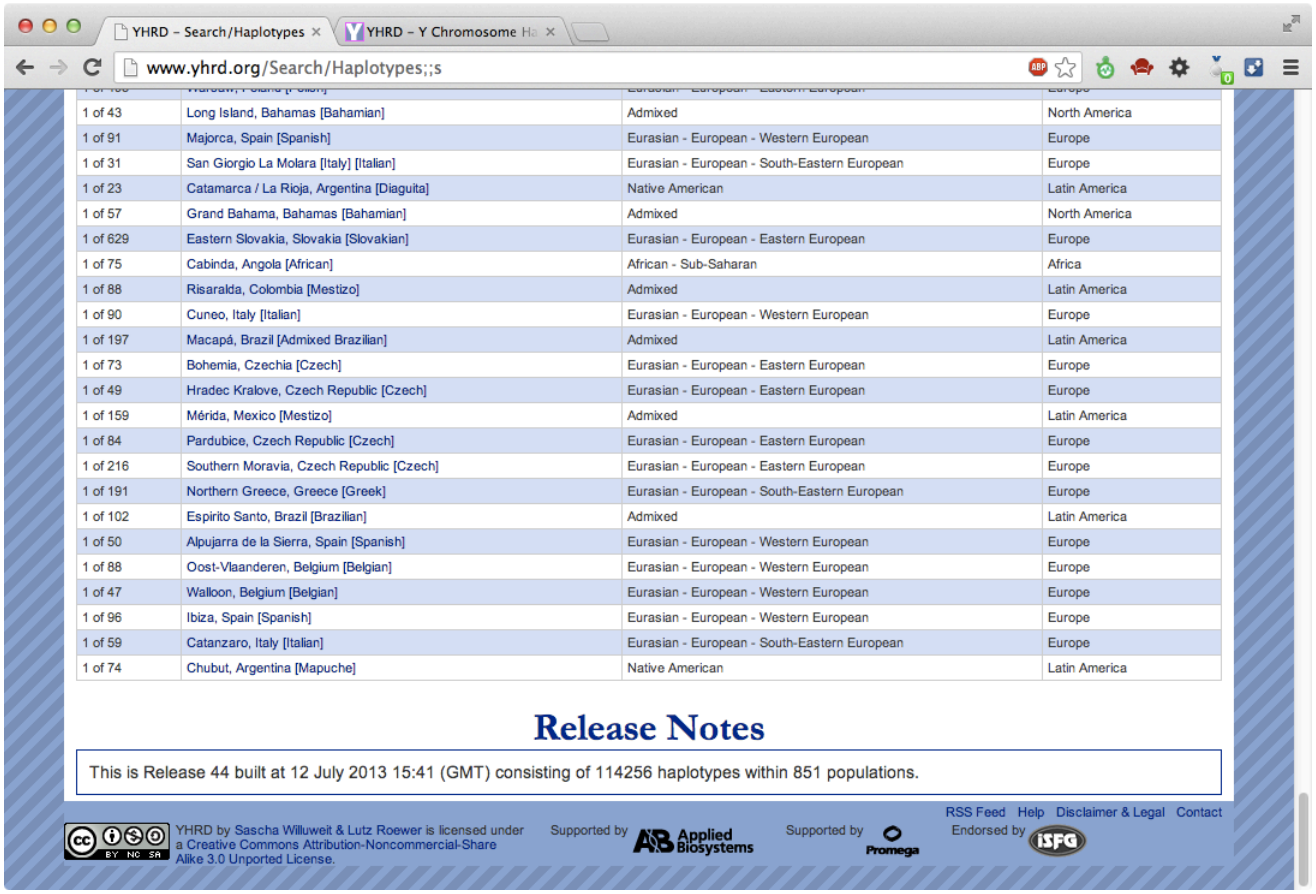


Figure 27: Search Haplotypes: Release notes

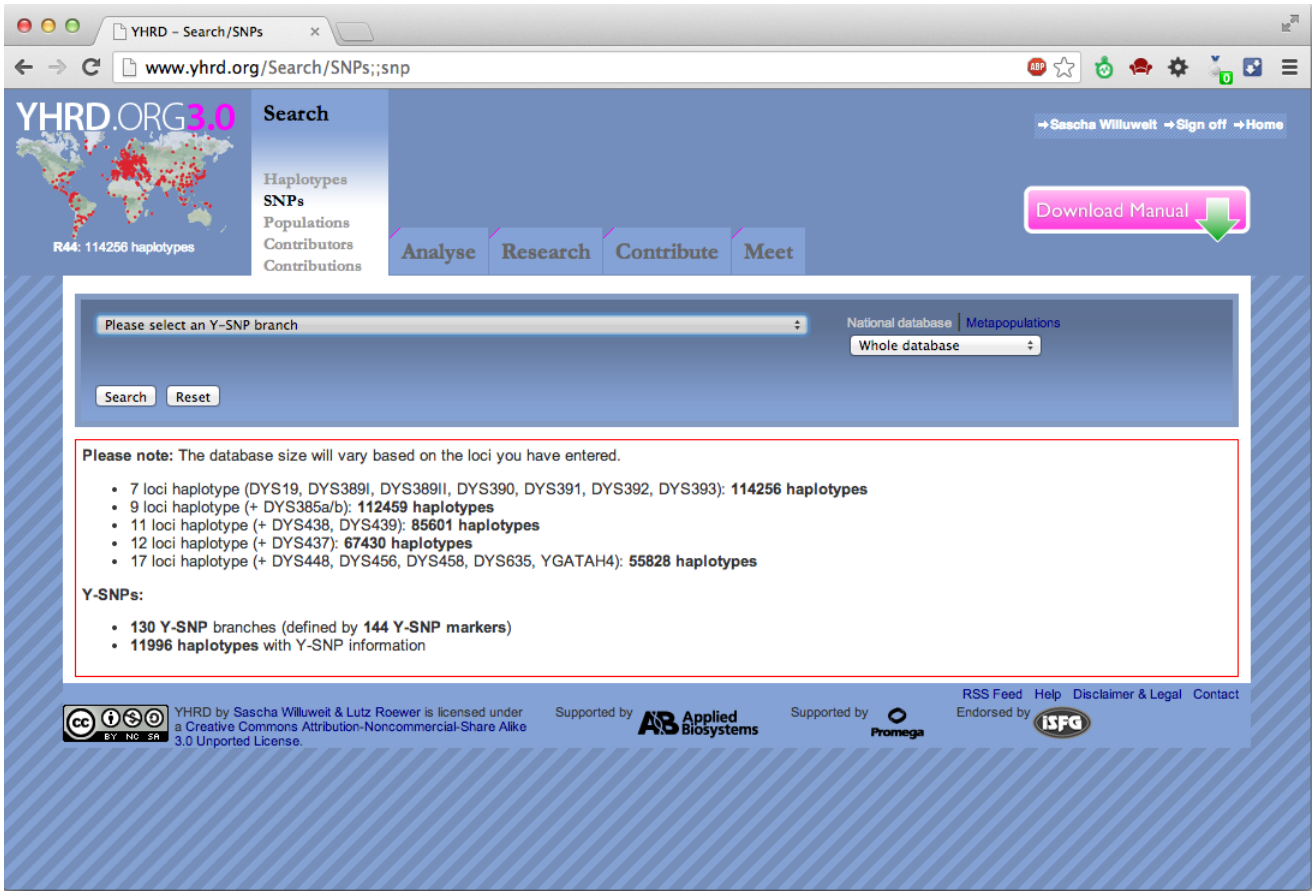


Figure 28: Search SNPs: Start

3.1.2 By SNPs – `http://www.yhrd.org/Search/SNPs`

The SNP Search (see Figure 28) option allows to select the Y-SNP / Haplogroup (see Figure 29) to receive statistics (analogous to Haplotypes search, see Figures 30, 31 and 32) and all STR haplotypes (see Figure 33) which were typed for the respective SNPs and fall in a certain branch (see Figures 28, 28). All Y-SNP branches and markers are linked to their corresponding fact sheet (see Figures 30 and 31).
A list of invalid Y-SNP markers can be found at Bad Y-SNPs Page[7] (see Figure 36).

[7] See `http://www.yhrd.org/Analyse/BYSNP`

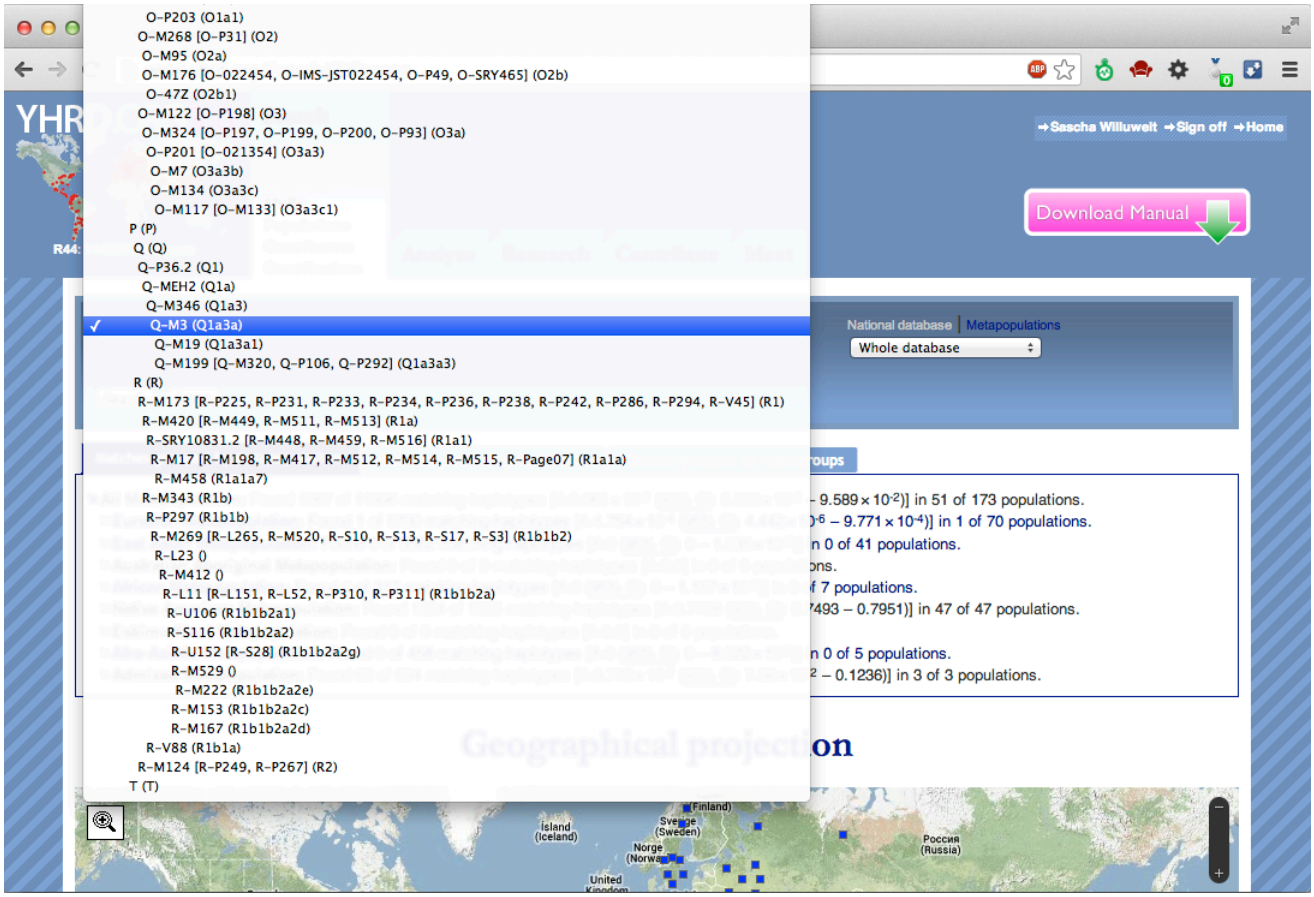


Figure 29: Search SNPs: Select a Y-SNP branch

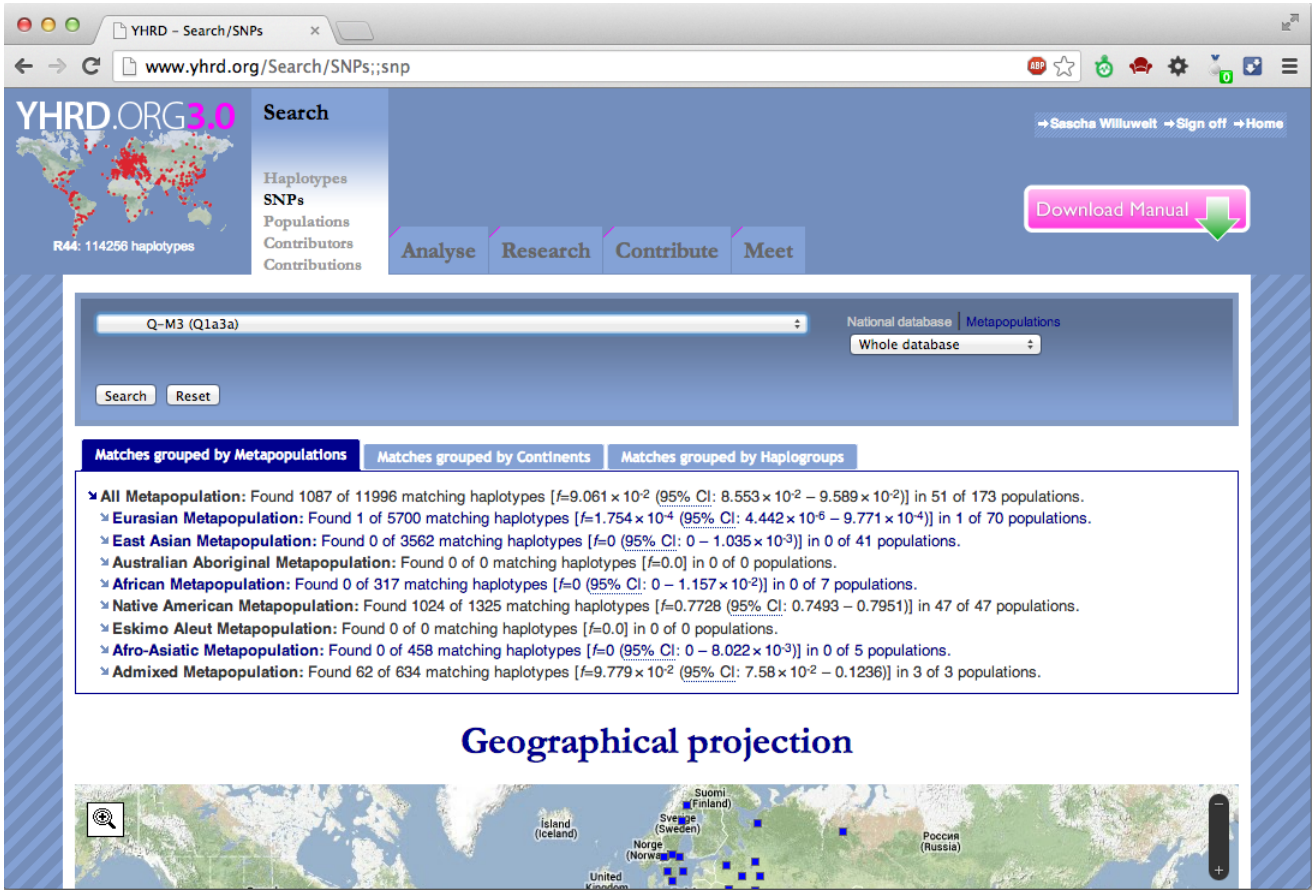


Figure 30: Search SNPs: Results by Metapopulation

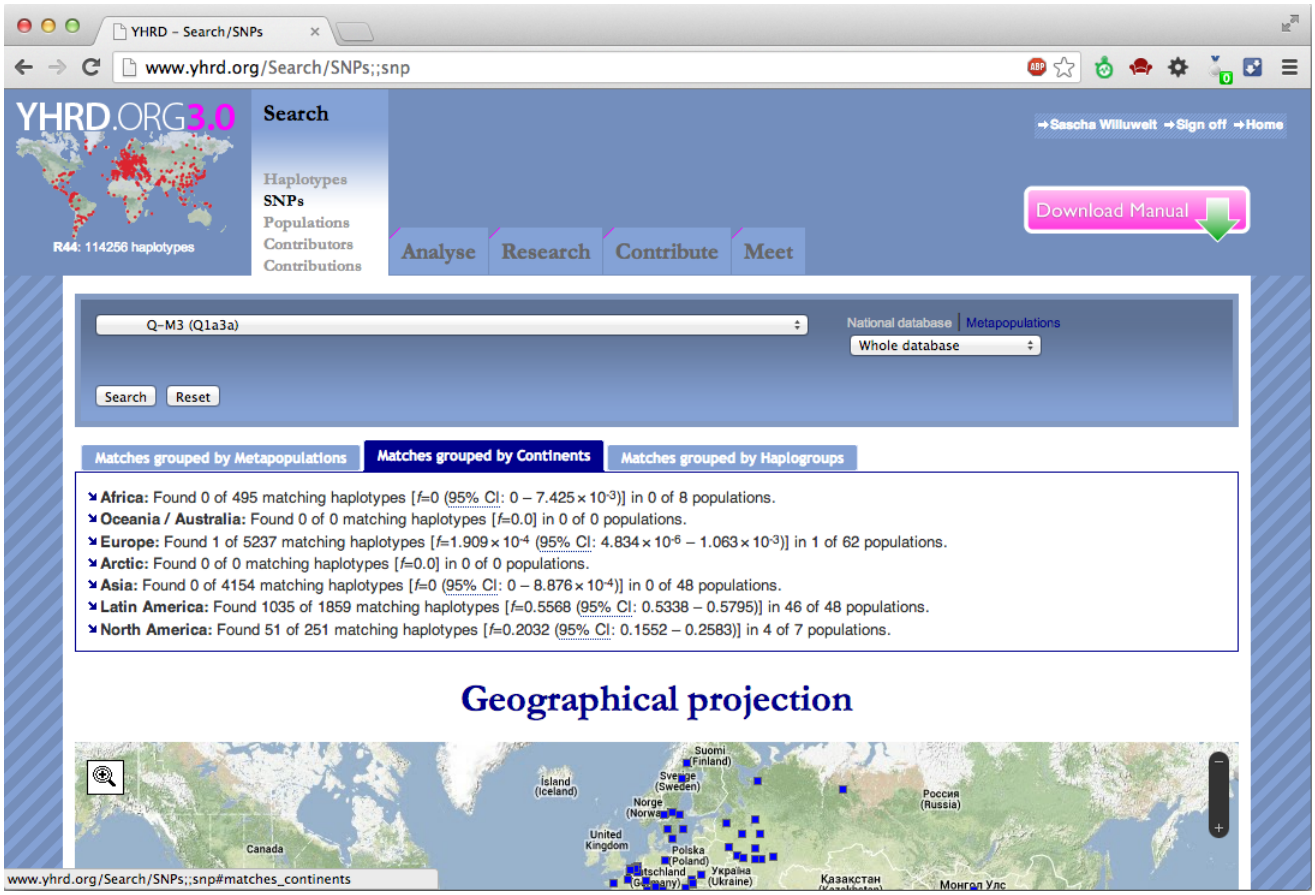


Figure 31: Search SNPs: Results by Continent

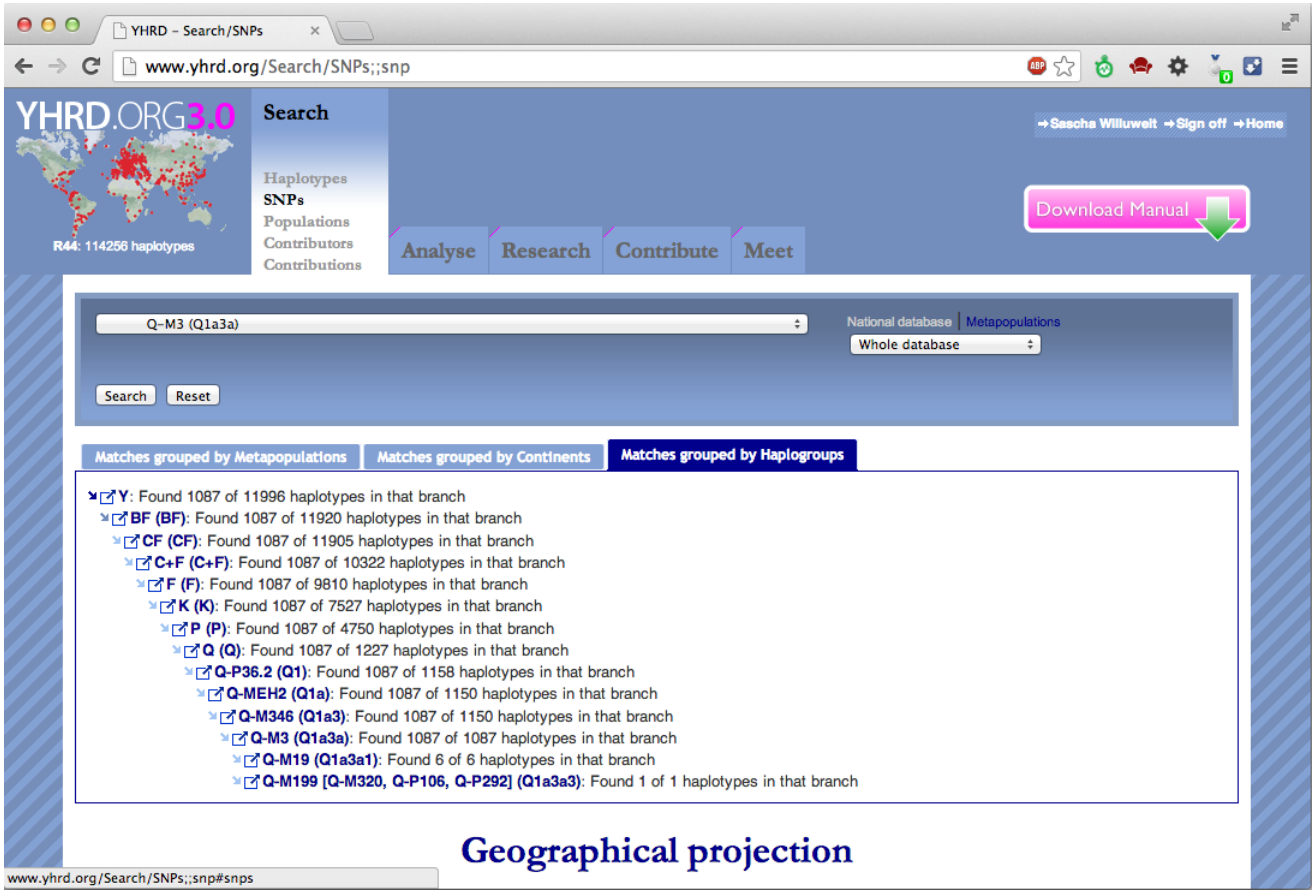
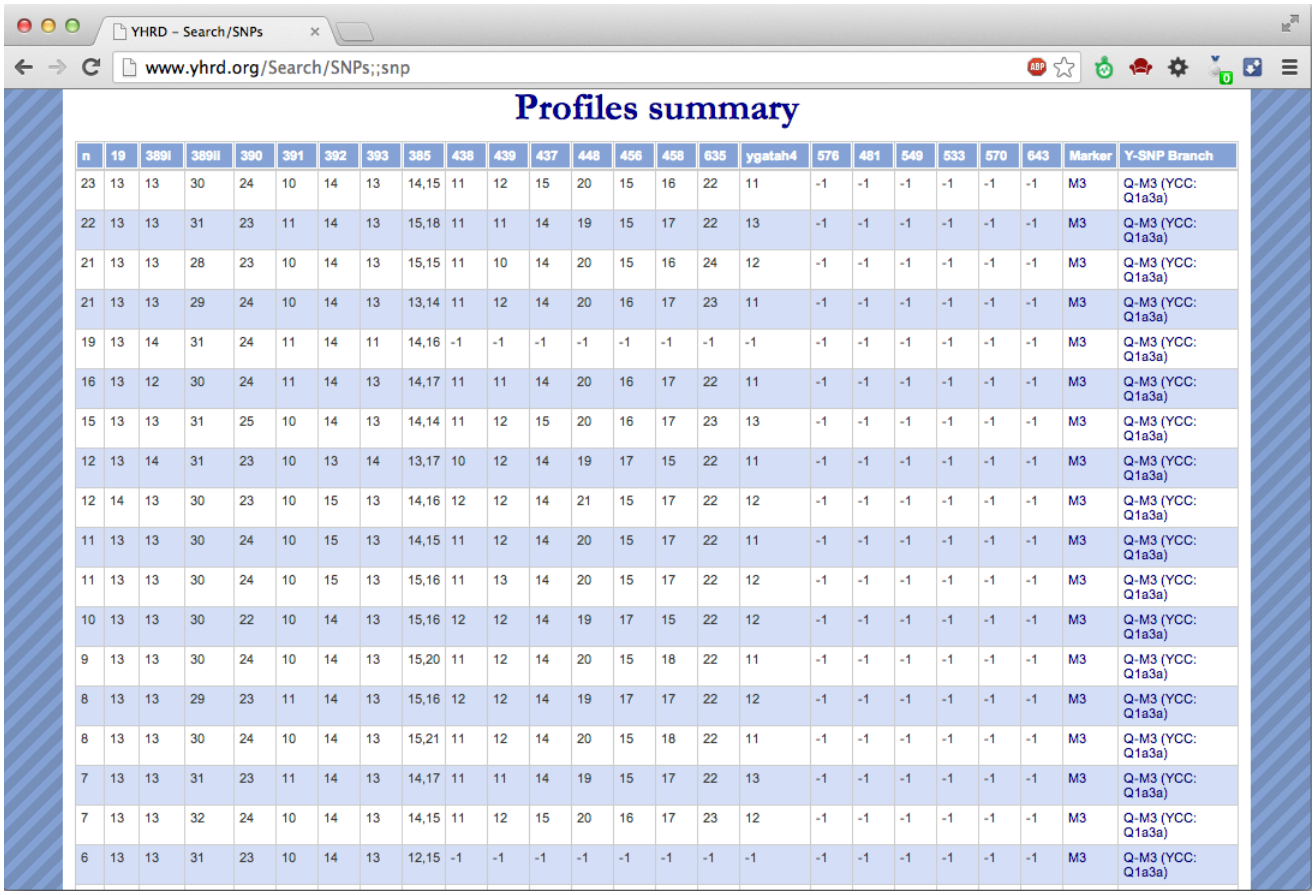


Figure 32: Search SNPs: Results by Haplogroup



Profiles summary

n	19	389I	389II	390	391	392	393	385	438	439	437	448	456	458	635	ygatah4	576	481	549	533	570	643	Marker	Y-SNP Branch
23	13	13	30	24	10	14	13	14,15	11	12	15	20	15	16	22	11	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
22	13	13	31	23	11	14	13	15,18	11	11	14	19	15	17	22	13	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
21	13	13	28	23	10	14	13	15,15	11	10	14	20	15	16	24	12	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
21	13	13	29	24	10	14	13	13,14	11	12	14	20	16	17	23	11	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
19	13	14	31	24	11	14	11	14,16	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
16	13	12	30	24	11	14	13	14,17	11	11	14	20	16	17	22	11	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
15	13	13	31	25	10	14	13	14,14	11	12	15	20	16	17	23	13	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
12	13	14	31	23	10	13	14	13,17	10	12	14	19	17	15	22	11	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
12	14	13	30	23	10	15	13	14,16	12	12	14	21	15	17	22	12	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
11	13	13	30	24	10	15	13	14,15	11	12	14	20	15	17	22	11	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
11	13	13	30	24	10	15	13	15,16	11	13	14	20	15	17	22	12	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
10	13	13	30	22	10	14	13	15,16	12	12	14	19	17	15	22	12	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
9	13	13	30	24	10	14	13	15,20	11	12	14	20	15	18	22	11	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
8	13	13	29	23	11	14	13	15,16	12	12	14	19	17	17	22	12	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
8	13	13	30	24	10	14	13	15,21	11	12	14	20	15	18	22	11	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
7	13	13	31	23	11	14	13	14,17	11	11	14	19	15	17	22	13	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
7	13	13	32	24	10	14	13	14,15	11	12	15	20	16	17	23	12	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)
6	13	13	31	23	10	14	13	12,15	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	M3	Q-M3 (YCC: Q1a3a)

Figure 33: Search SNPs: Resulting Haplotypes

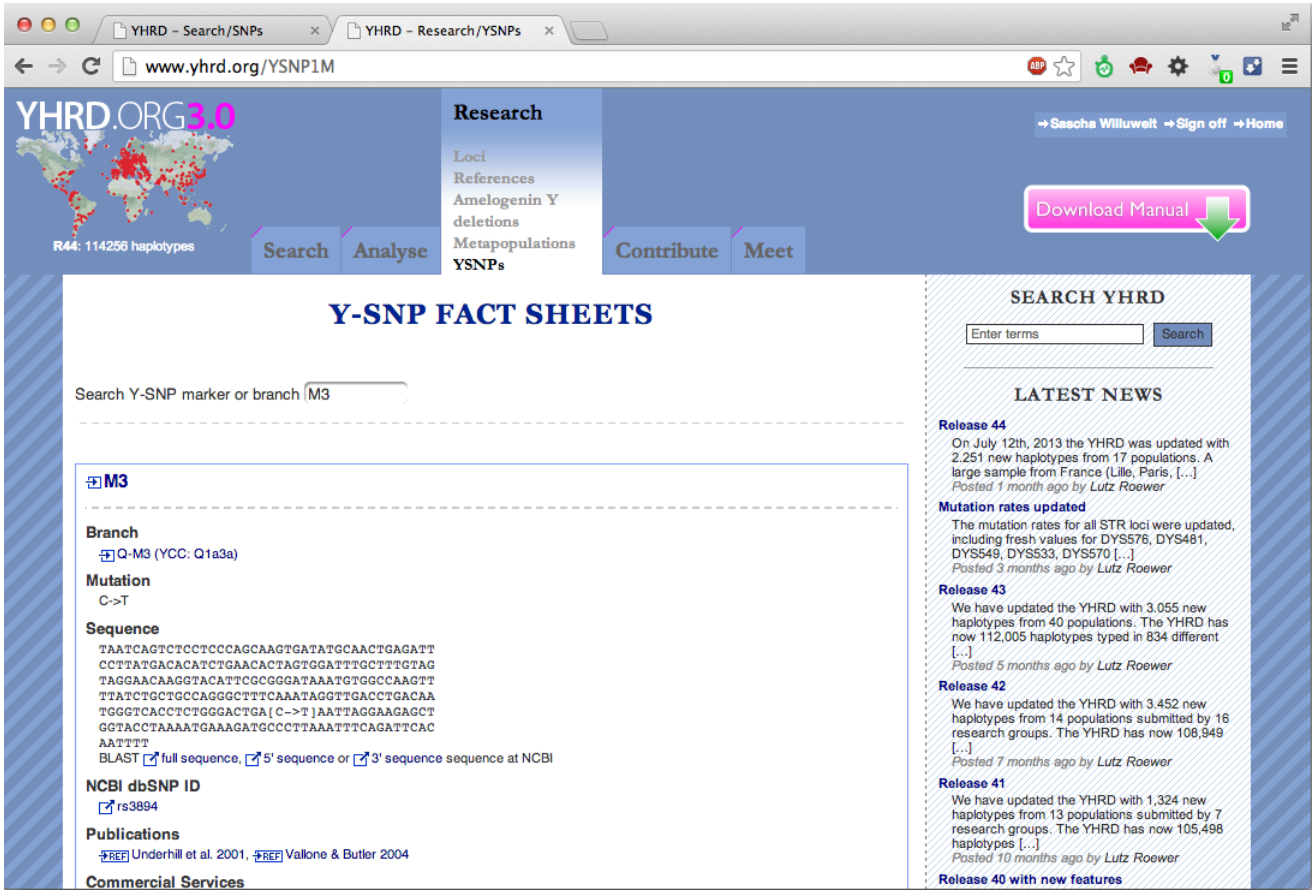


Figure 34: Search SNPs: Linked Y-SNP fact sheet (branch)

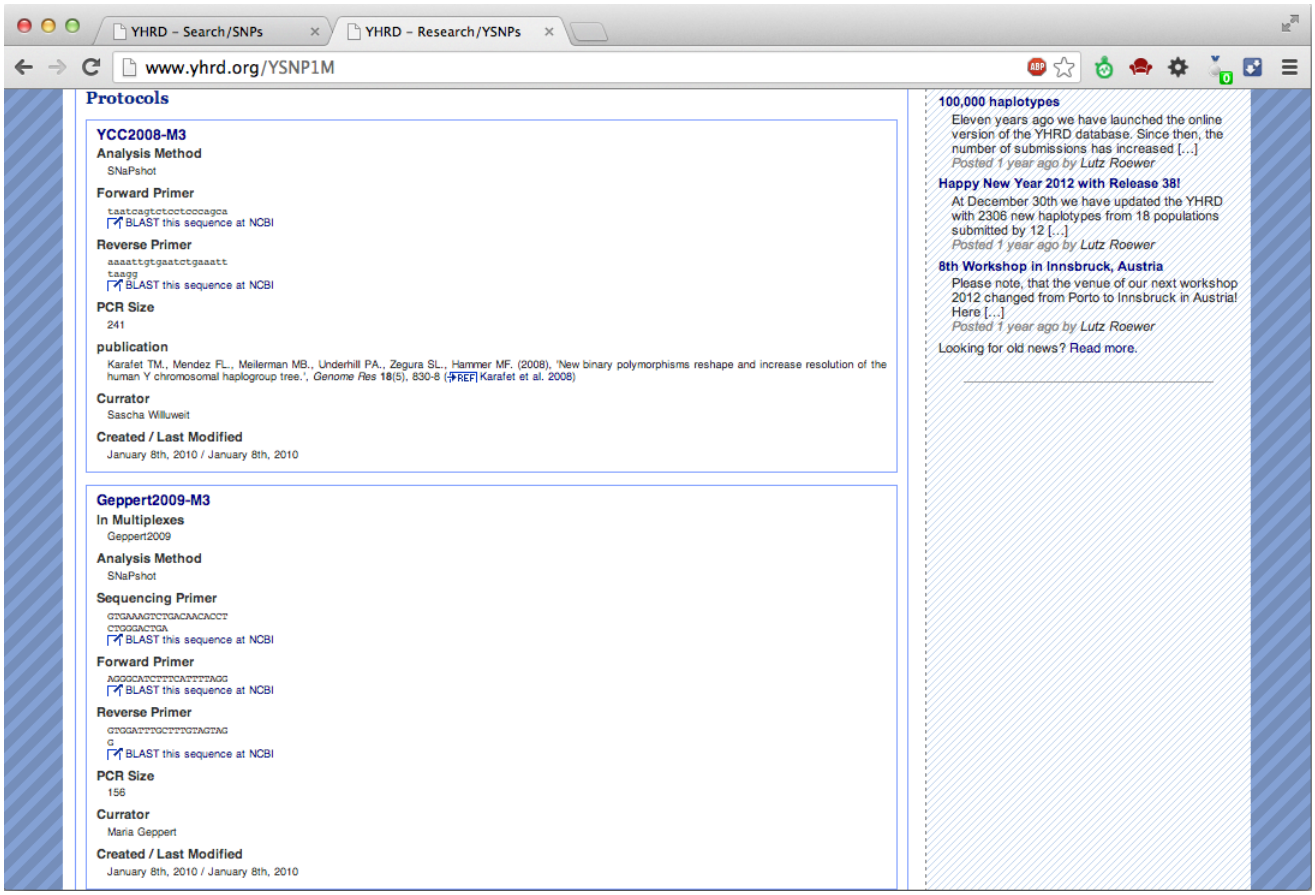


Figure 35: Search SNPs: Linked Y-SNP fact sheet (marker)

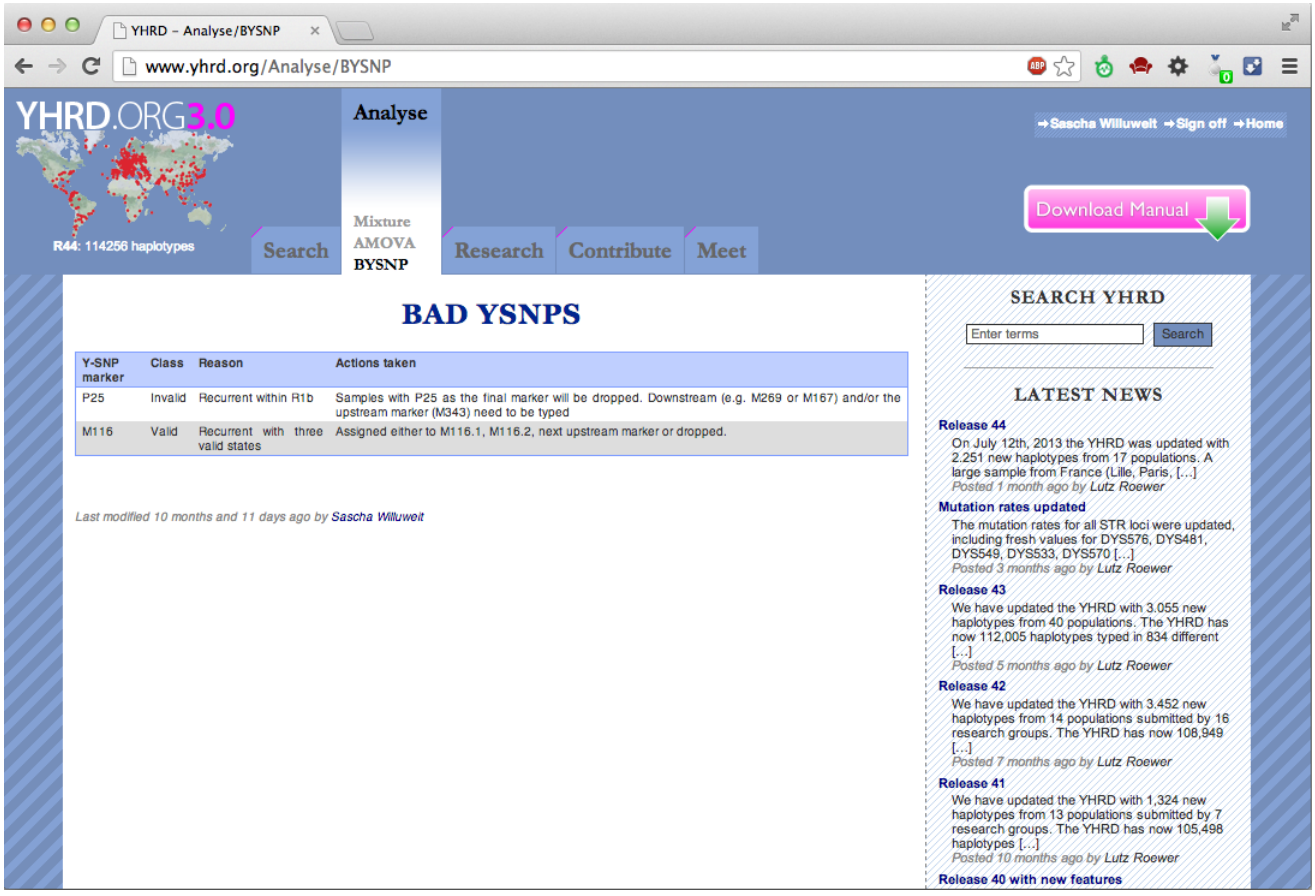


Figure 36: BYSNP: A list of invalid Y-SNP markers

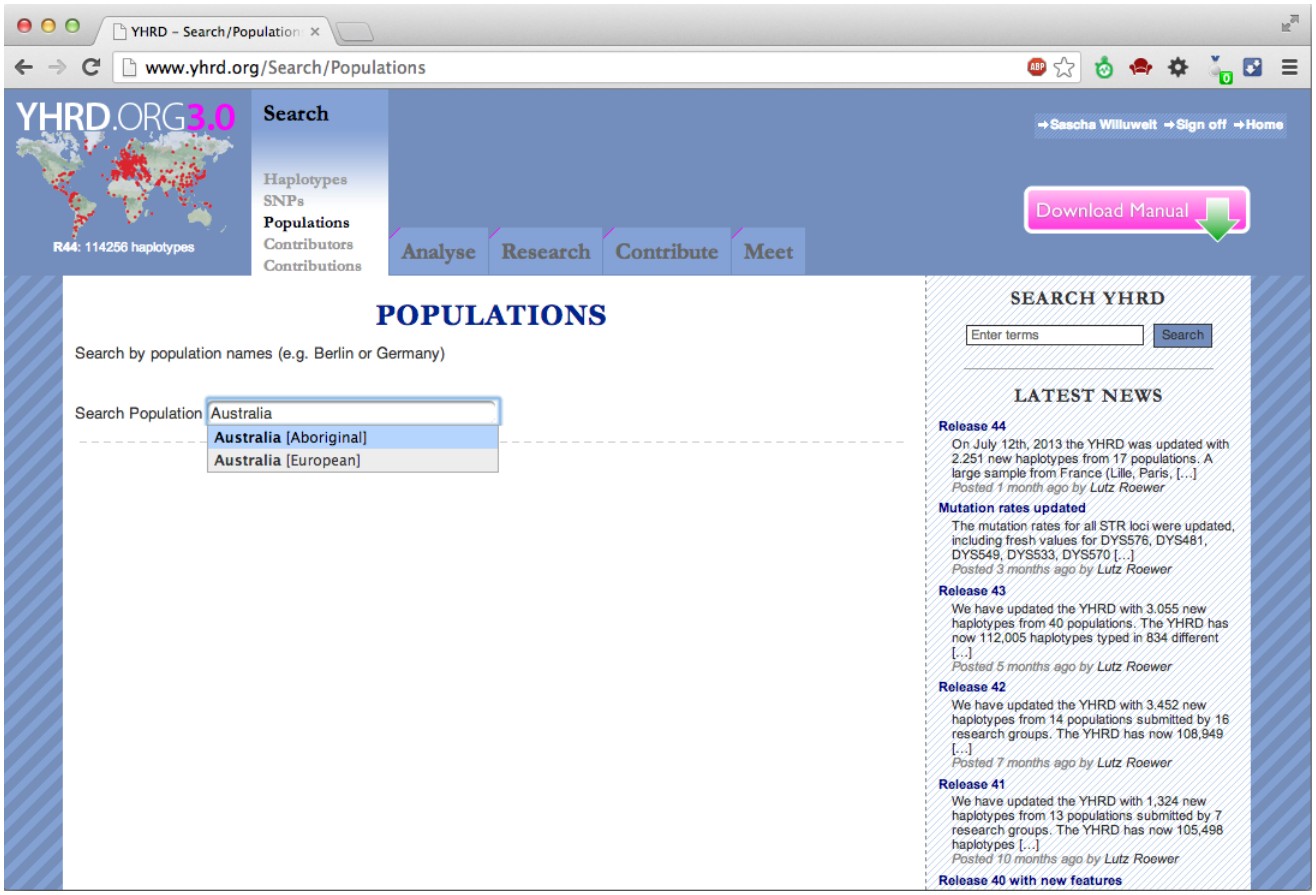


Figure 37: Search Populations: Autocomplete feature

3.1.3 By Population – `http://www.yhrd.org/Search/Population`

Search for a population study by its name or by a part of its name. As a result a document with essential information on the sampling project is given, including name, coordinates, Geoposition (movable and scalable Google maps functionality), assigned Metapopulation, sample size, contributor information, most frequent Haplotypes and Haplotypes with irregular alleles ("Mutations"). Headers can be clicked to change the sorting order and all entries in blue are linked to the respective pages (see Figures 37, 38, 39 and 40).

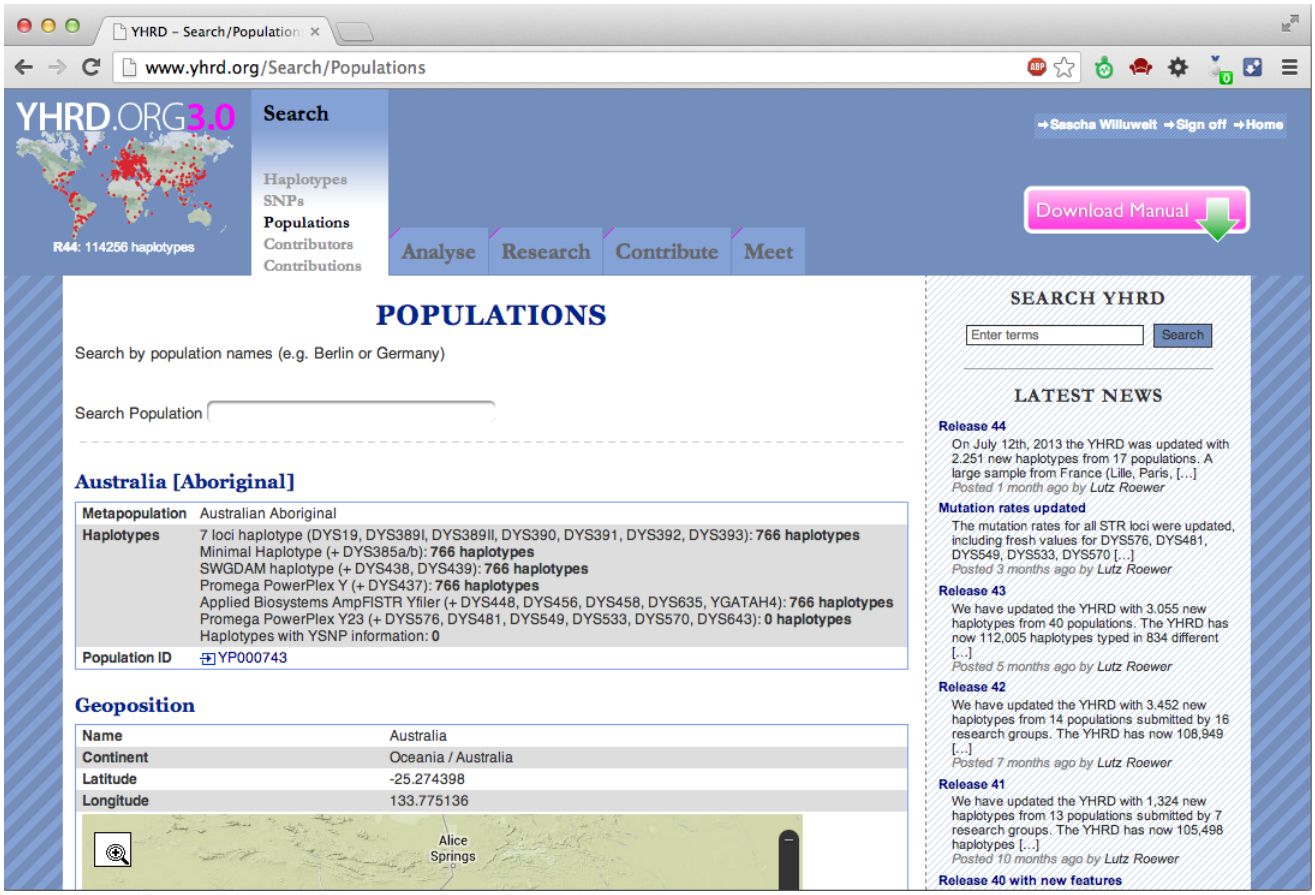


Figure 38: Search Populations: Result: General information

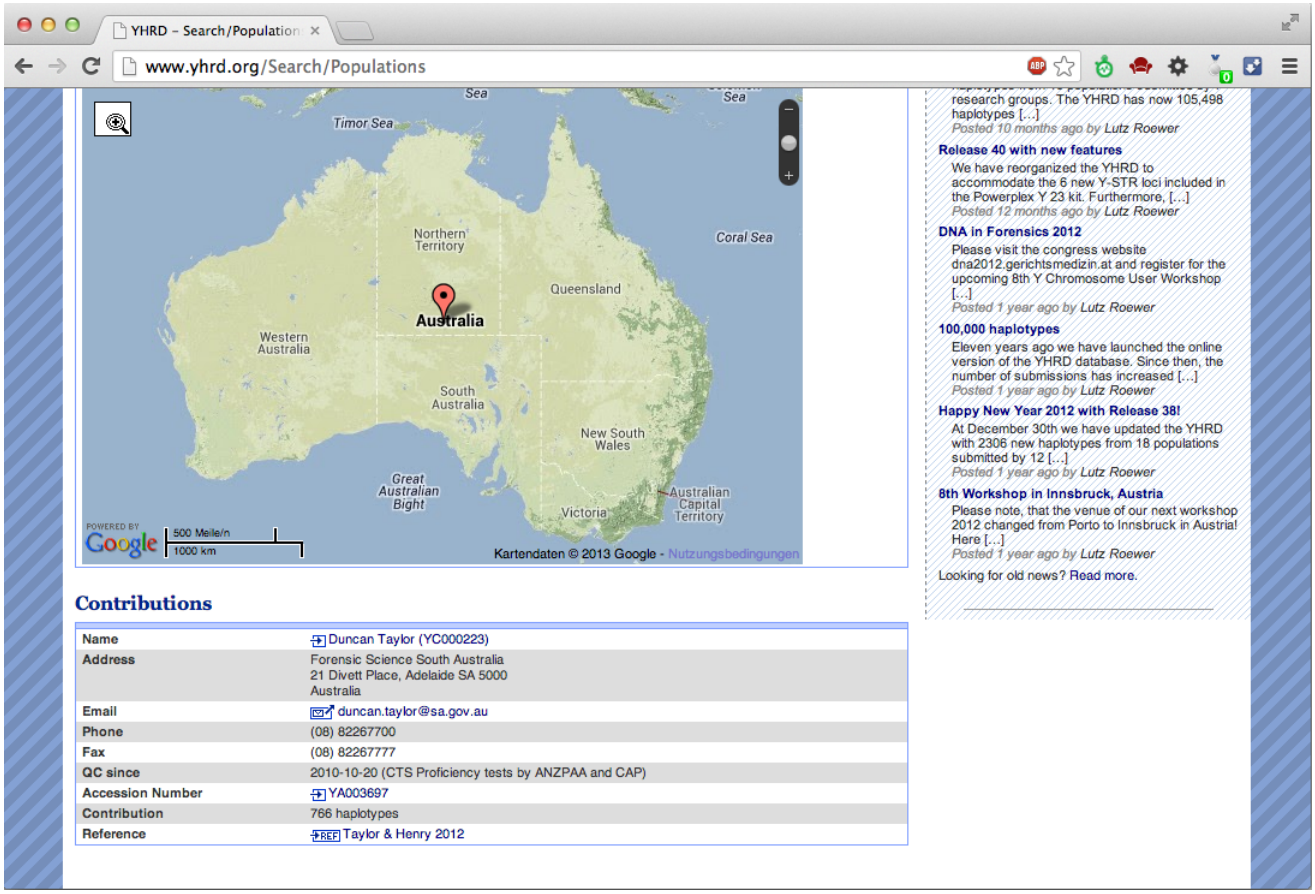
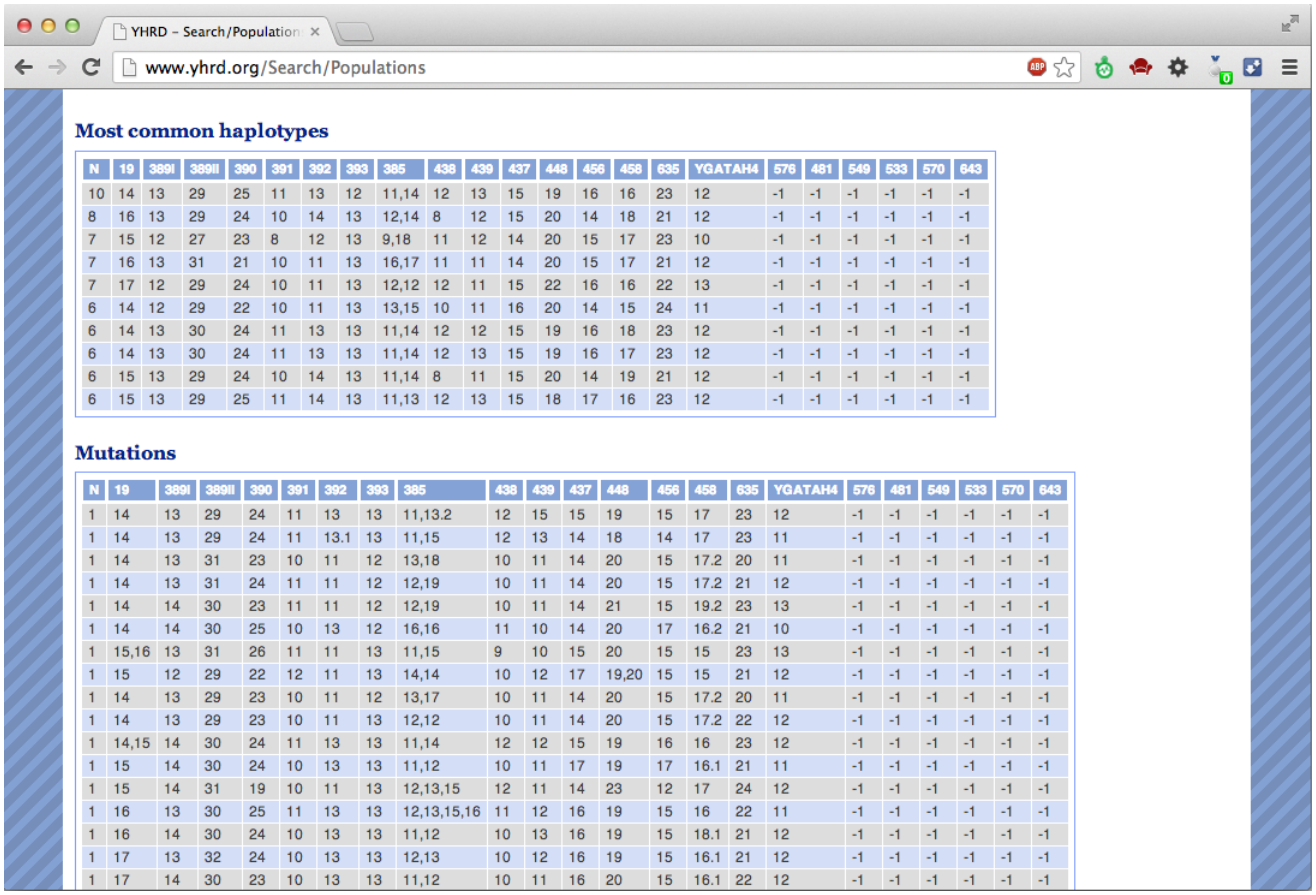


Figure 39: Search Populations: Result: Map and contributor



N	19	389I	389II	390	391	392	393	385	438	439	437	448	456	458	635	YGATAH4	576	481	549	533	570	643
10	14	13	29	25	11	13	12	11,14	12	13	15	19	16	16	23	12	-1	-1	-1	-1	-1	-1
8	16	13	29	24	10	14	13	12,14	8	12	15	20	14	18	21	12	-1	-1	-1	-1	-1	-1
7	15	12	27	23	8	12	13	9,18	11	12	14	20	15	17	23	10	-1	-1	-1	-1	-1	-1
7	16	13	31	21	10	11	13	16,17	11	11	14	20	15	17	21	12	-1	-1	-1	-1	-1	-1
7	17	12	29	24	10	11	13	12,12	12	11	15	22	16	16	22	13	-1	-1	-1	-1	-1	-1
6	14	12	29	22	10	11	13	13,15	10	11	16	20	14	15	24	11	-1	-1	-1	-1	-1	-1
6	14	13	30	24	11	13	13	11,14	12	12	15	19	16	18	23	12	-1	-1	-1	-1	-1	-1
6	14	13	30	24	11	13	13	11,14	12	13	15	19	16	17	23	12	-1	-1	-1	-1	-1	-1
6	15	13	29	24	10	14	13	11,14	8	11	15	20	14	19	21	12	-1	-1	-1	-1	-1	-1
6	15	13	29	25	11	14	13	11,13	12	13	15	18	17	16	23	12	-1	-1	-1	-1	-1	-1

N	19	389I	389II	390	391	392	393	385	438	439	437	448	456	458	635	YGATAH4	576	481	549	533	570	643
1	14	13	29	24	11	13	13	11,13.2	12	15	15	19	15	17	23	12	-1	-1	-1	-1	-1	-1
1	14	13	29	24	11	13.1	13	11,15	12	13	14	18	14	17	23	11	-1	-1	-1	-1	-1	-1
1	14	13	31	23	10	11	12	13,18	10	11	14	20	15	17.2	20	11	-1	-1	-1	-1	-1	-1
1	14	13	31	24	11	11	12	12,19	10	11	14	20	15	17.2	21	12	-1	-1	-1	-1	-1	-1
1	14	14	30	23	11	11	12	12,19	10	11	14	21	15	19.2	23	13	-1	-1	-1	-1	-1	-1
1	14	14	30	25	10	13	12	16,16	11	10	14	20	17	16.2	21	10	-1	-1	-1	-1	-1	-1
1	15,16	13	31	26	11	11	13	11,15	9	10	15	20	15	15	23	13	-1	-1	-1	-1	-1	-1
1	15	12	29	22	12	11	13	14,14	10	12	17	19,20	15	15	21	12	-1	-1	-1	-1	-1	-1
1	14	13	29	23	10	11	12	13,17	10	11	14	20	15	17.2	20	11	-1	-1	-1	-1	-1	-1
1	14	13	29	23	10	11	13	12,12	10	11	14	20	15	17.2	22	12	-1	-1	-1	-1	-1	-1
1	14,15	14	30	24	11	13	13	11,14	12	12	15	19	16	16	23	12	-1	-1	-1	-1	-1	-1
1	15	14	30	24	10	13	13	11,12	10	11	17	19	17	16.1	21	11	-1	-1	-1	-1	-1	-1
1	15	14	31	19	10	11	13	12,13,15	12	11	14	23	12	17	24	12	-1	-1	-1	-1	-1	-1
1	16	13	30	25	11	13	13	12,13,15,16	11	12	16	19	15	16	22	11	-1	-1	-1	-1	-1	-1
1	16	14	30	24	10	13	13	11,12	10	13	16	19	15	18.1	21	12	-1	-1	-1	-1	-1	-1
1	17	13	32	24	10	13	13	12,13	10	12	16	19	15	16.1	21	12	-1	-1	-1	-1	-1	-1
1	17	14	30	23	10	13	13	11,12	10	11	16	20	15	16.1	22	12	-1	-1	-1	-1	-1	-1

Figure 40: Search Populations: Result: Most common Haplotypes and Mutations

3.1.4 By Contributor – `http://www.yhrd.org/Search/Contributor`

Search for a colleague and database contributor or for one of his group by name or part of the name (see Figure 41). You'll receive information in form of a "Business Card" including address, email contact and telephone number (if submitted) as well as the date of the certificate on the participation in the Quality Control test (see Section 3.4).

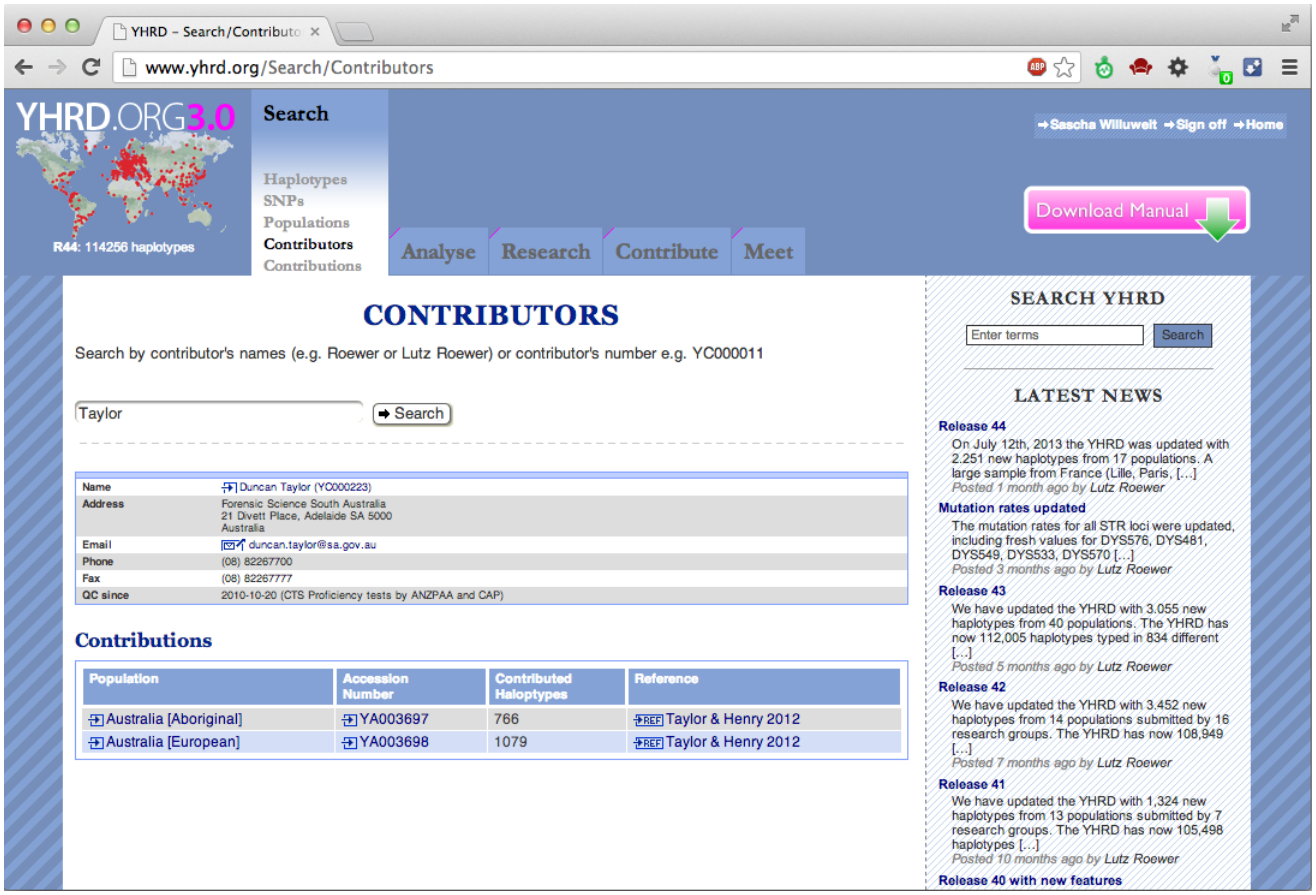


Figure 41: Search Contributor

3.1.5 By Contribution (Accession Number) – `http://www.yhrd.org/Search/Contribution`

It is mandatory for any new contributor to pass a "Quality test", which means a correct typing of 5 blind DNA samples for the Y-STR markers which he is going to submit; the results will be evaluated and certified by the curators. This test is required only once, future contributions or updates of your samples will not require a new QC.
After passing the QC and issuing of a certificate population data can be submitted anytime. We, as the curatorial board perform a quality check of these data which includes (a) a plausibility check to find obvious typos and (b) a similarity / genetic distance check, which includes a comparison of your data with other closely related population samples from YHRD to find out whether your data deviate from the typical core Haplotypes of the region. Among other measures we use the AMOVA (Analysis of Molecular Variance) to test for genetic distances between your sample and reference samples (see Section 3.2.1). After this external quality and plausibility check we issue an YHRD accession number for each submitted population study of the contributing lab. It is also mandatory to include this number in articles submitted to journals (e.g. FSI:Genetics[8]) as a proof of external validation of the Haplotype data. The search for the Accession Number returns information on the respected population study and on the responsible author behind (see Figure 42).

[8] See `http://www.fsigenetics.com`

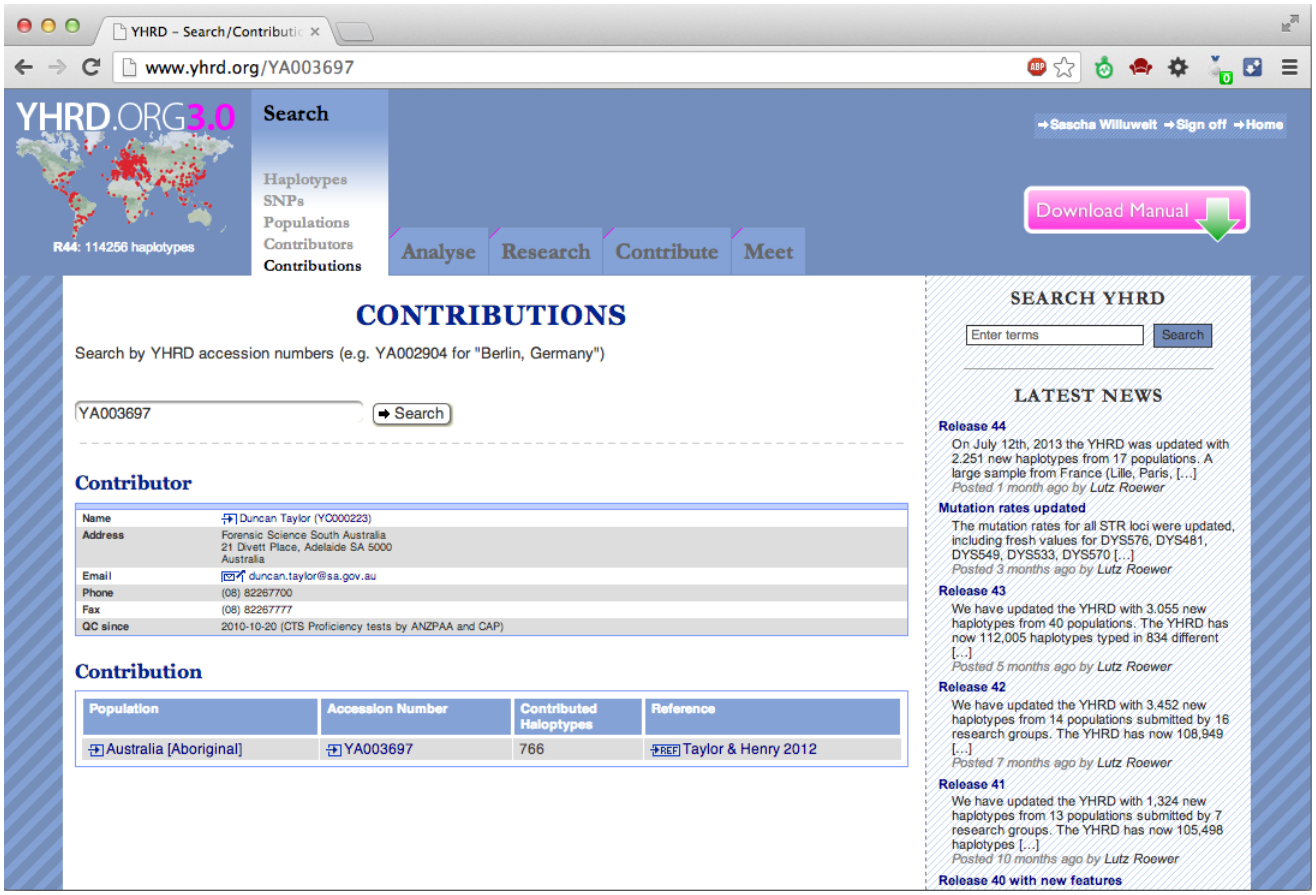


Figure 42: Search Contribution (Accession Number)

3.2 Analyse – `http://www.yhrd.org/Analyse`

3.2.1 AMOVA – `http://www.yhrd.org/Analyse/Online+AMOVA`

Analysis of Molecular Variance (AMOVA) is a method for analyzing population variation using molecular data, e.g. Y-STR Haplotypes (Roewer et al. (1996)). With AMOVA it is possible to evaluate and quantify the extent of differentiation between two or more population samples. AMOVA is implemented as an online tool in the YHRD and provides a way of estimating $\Phi_{ST}(R_{ST})$ and F_{ST} values. The online tool included in http://www.yhrd.org/Analyse accepts your Excel files and creates entry files from it. Attention: All entries highlighted in red will be ignored (e.g. a column ID or population name). So make clear that if you want to compare YFiler Haplotypes with reference studies including only minimal Haplotypes all additional loci are ignored. After you have submitted your entry file the program asks to confirm it and you can still do changes if necessary. As much as 9 reference populations selected from the YHRD as well as population sets can be added to the AMOVA analysis. The online calculation returns as a result a *.csv table with pairwise F_{ST} or $\Phi_{ST}(R_{ST})$ values plus p-values as a test for significance (10,000 permutations). In addition, an MDS plot is generated to illustrate the genetic distance between the analyzed populations graphically. The program shows the references for the selected population studies which facilitates the correct citation. The procedure step-by-step is explained at the first page (see Figure 43) and illustrated by Figures 44, 45, 46, 47, 48, 49, 50, 51, 52 and 53.

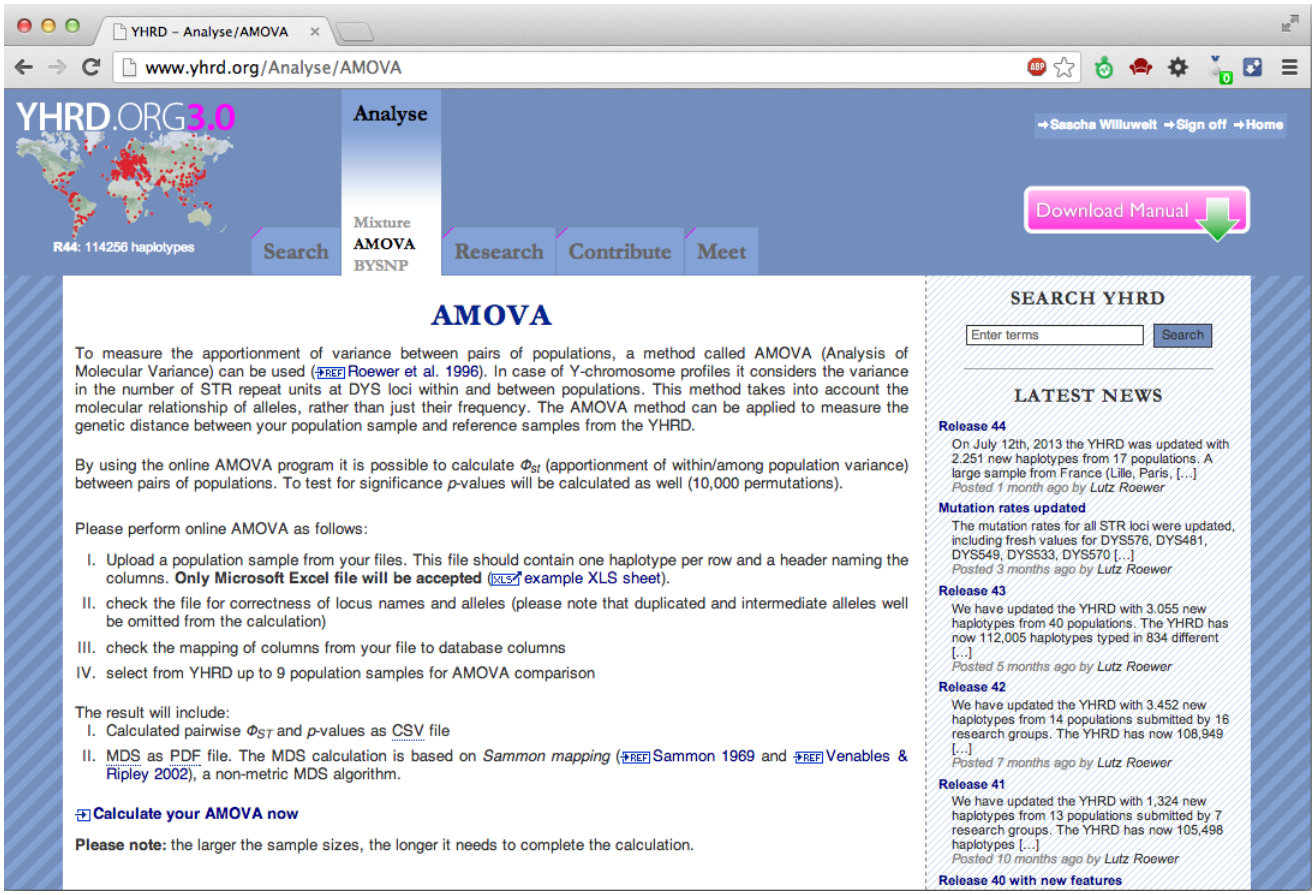


Figure 43: AMOVA: Guide and list of requirements

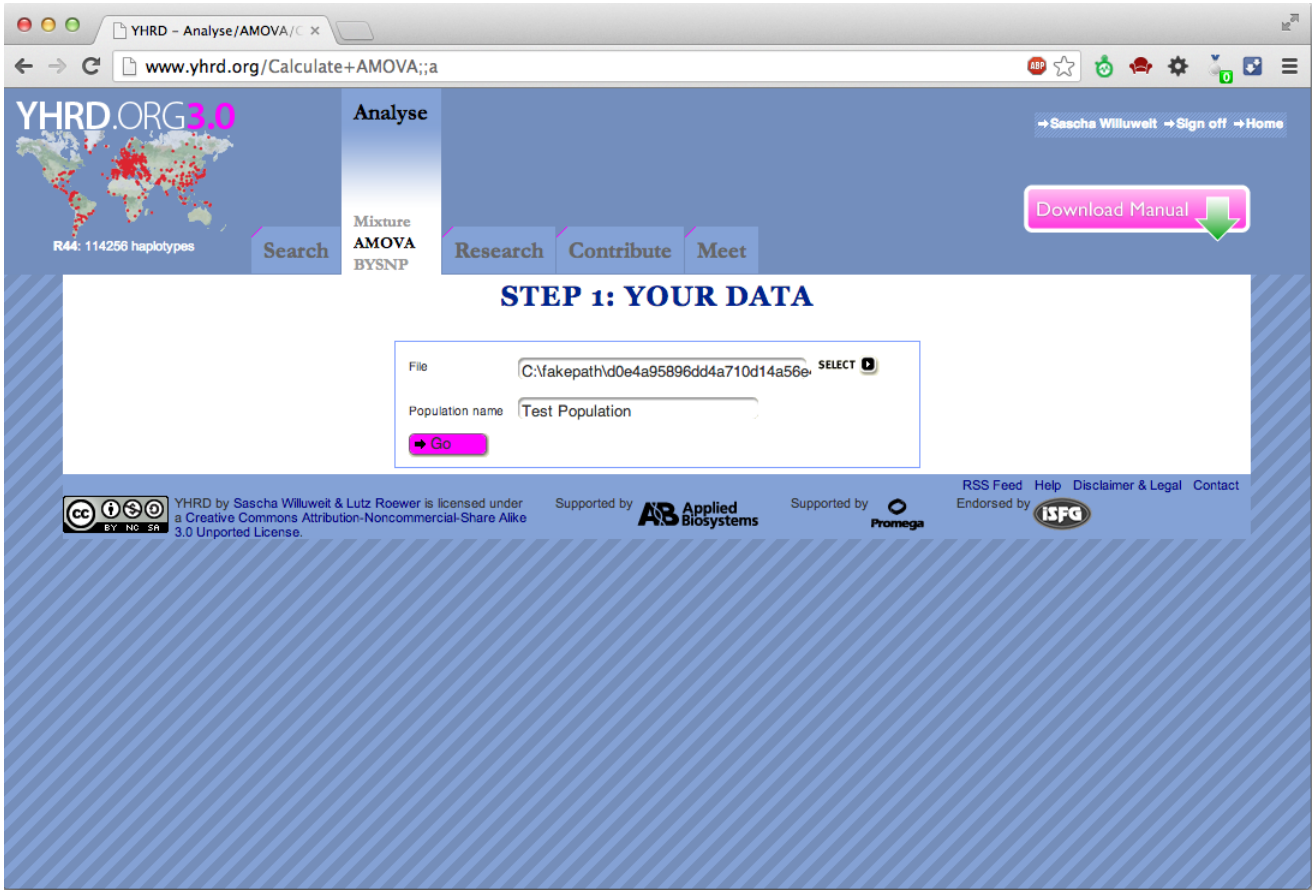


Figure 44: AMOVA: Step 1: Enter your data

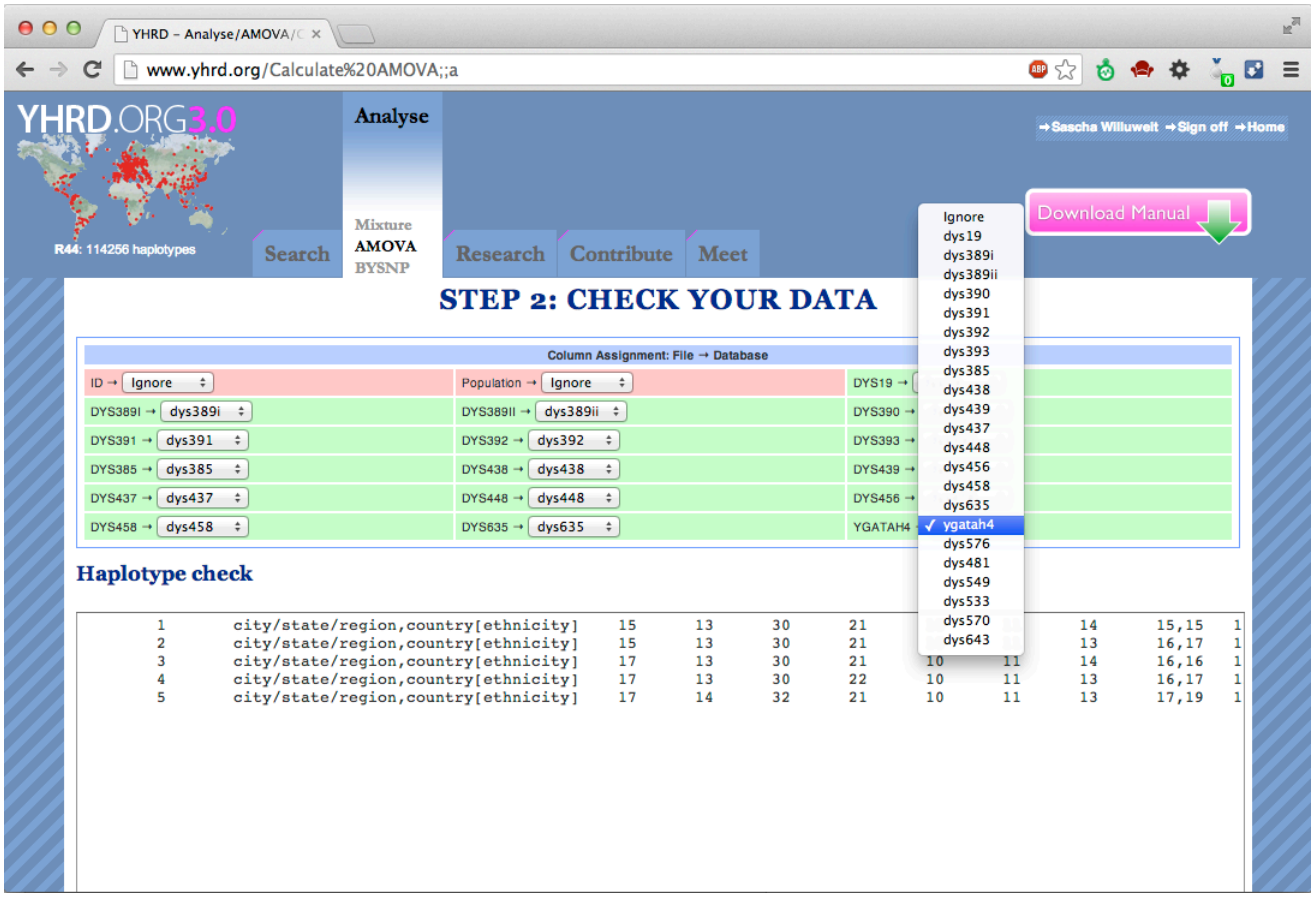


Figure 45: AMOVA: Step 2: Check your data: Validate locus assignment

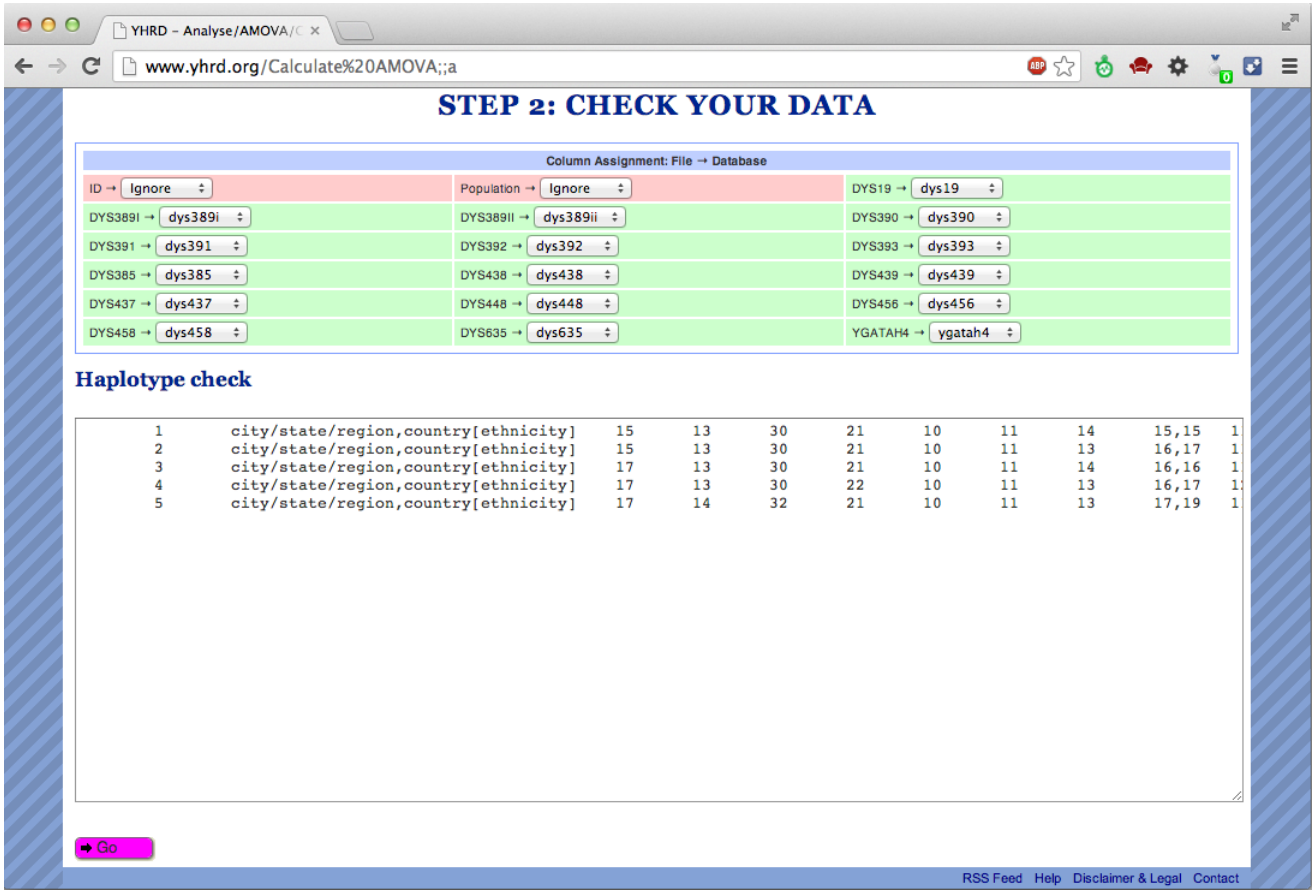


Figure 46: AMOVA: Step 2: Check your data

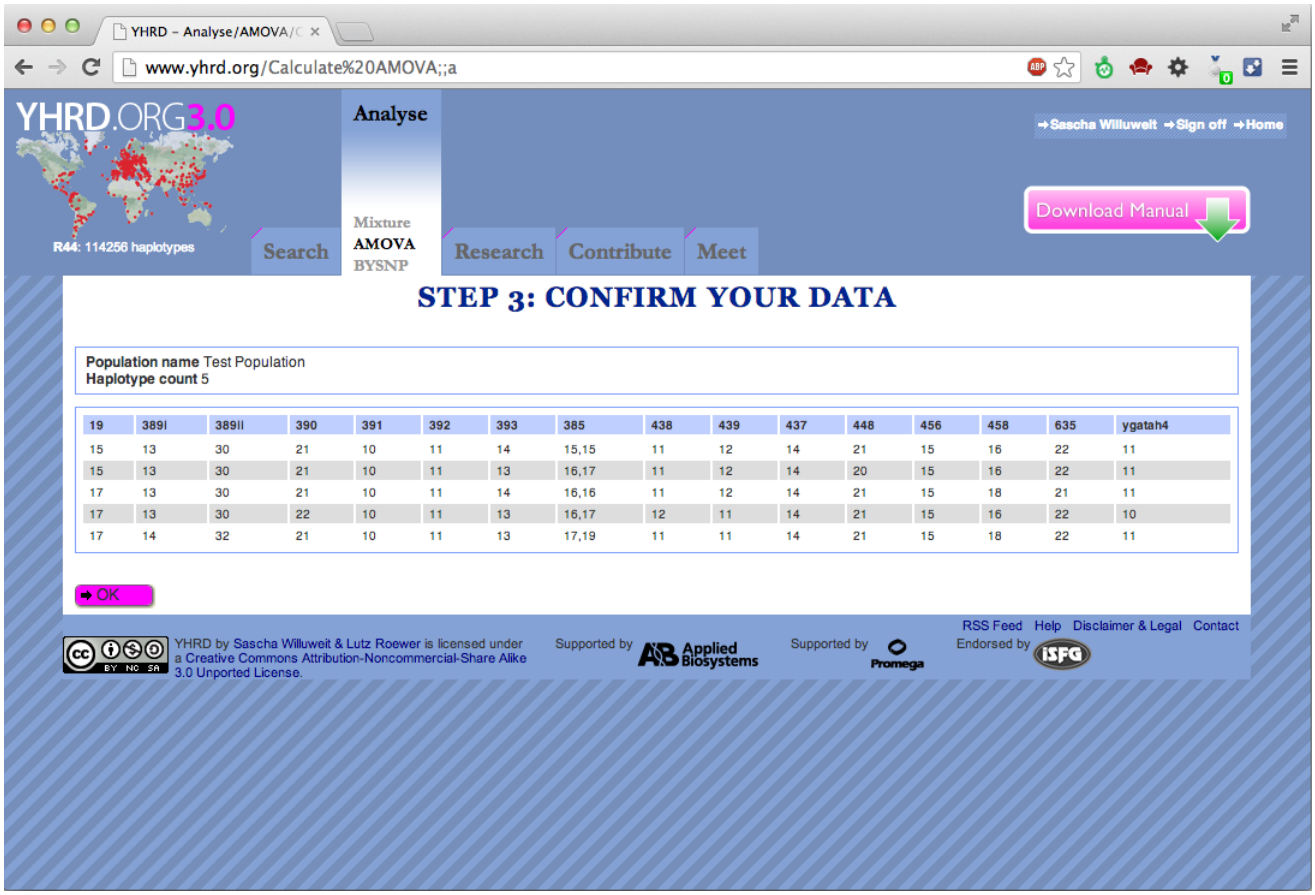


Figure 47: AMOVA: Step 3: Confirm your data

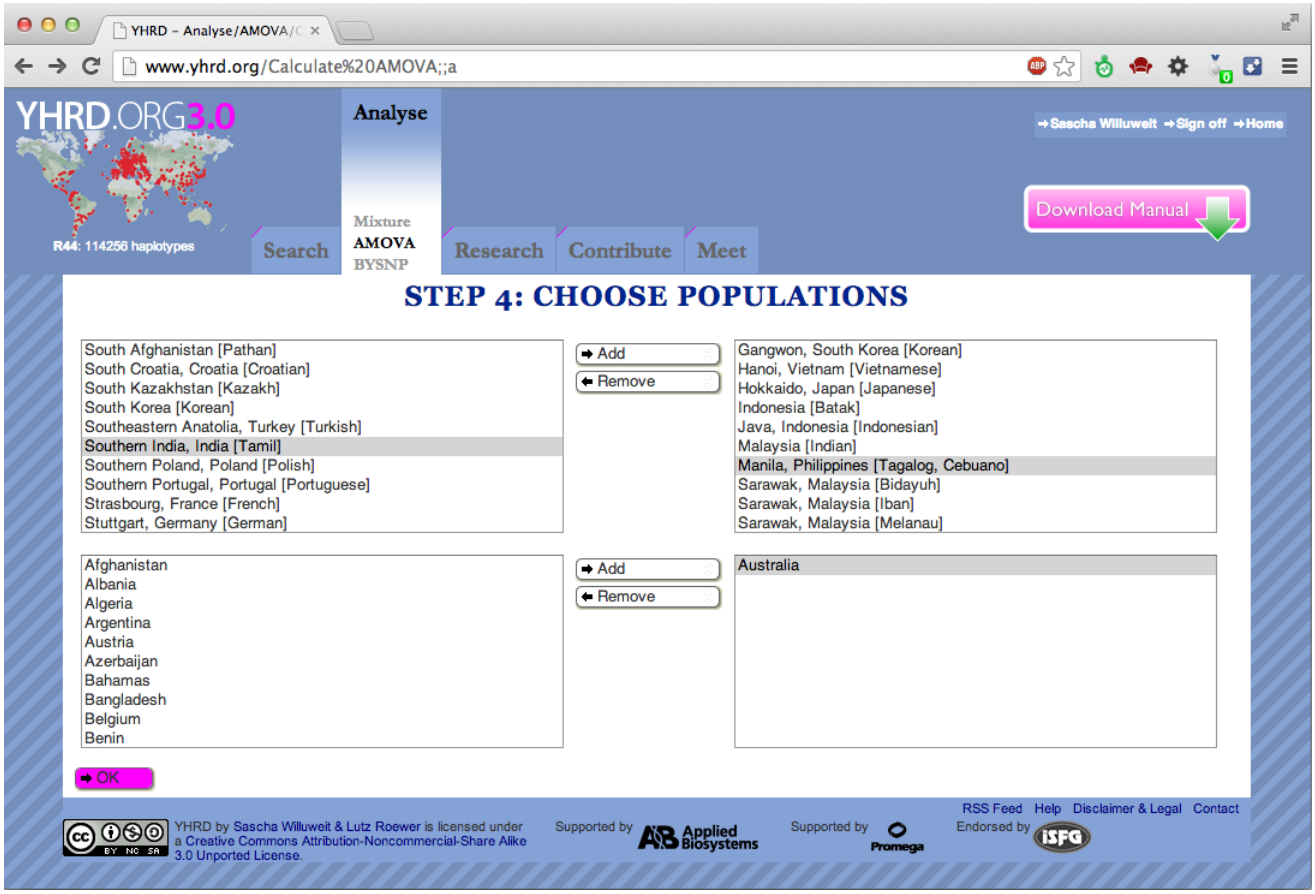


Figure 48: AMOVA: Step 4: Choose population and population sets

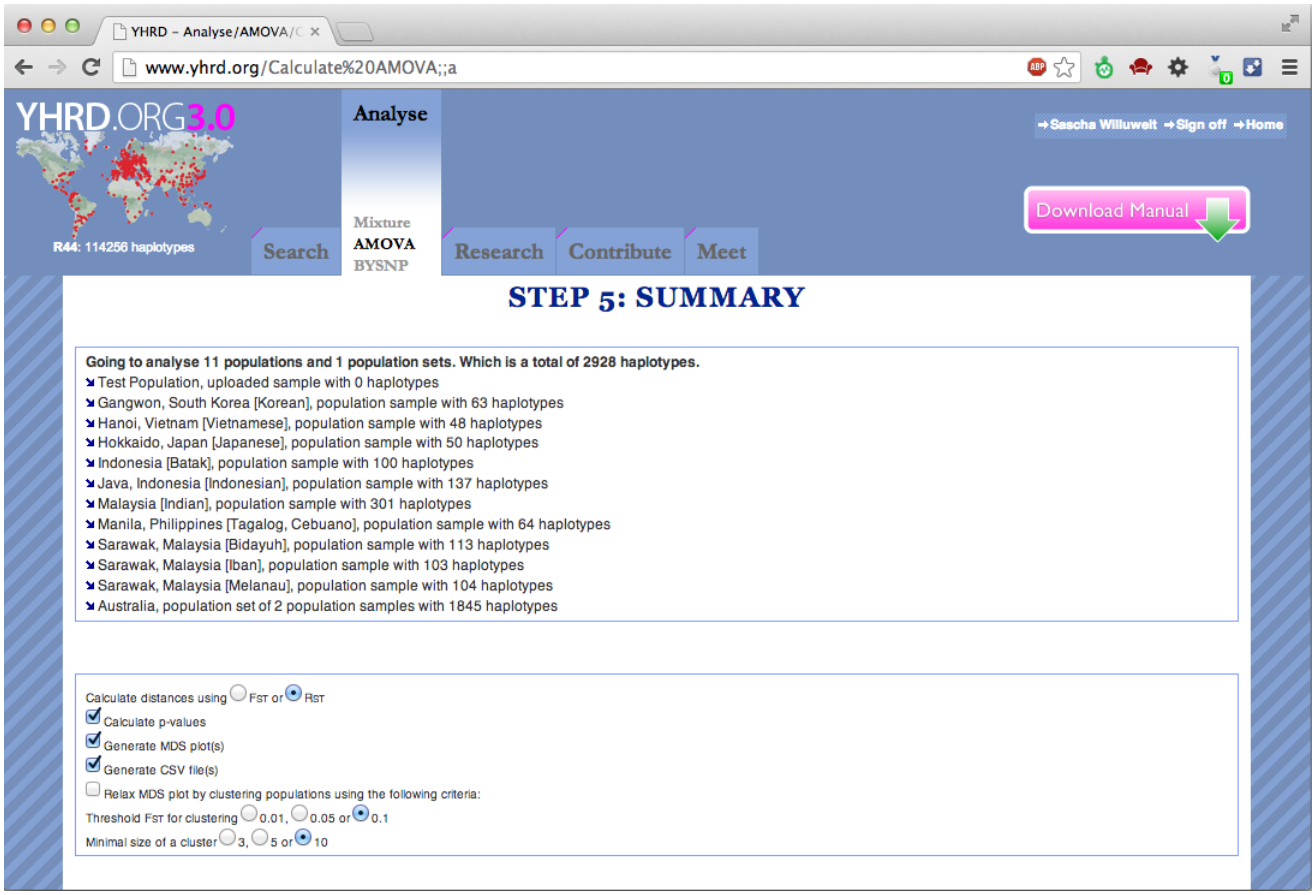


Figure 49: AMOVA: Step 5: Calculation summary

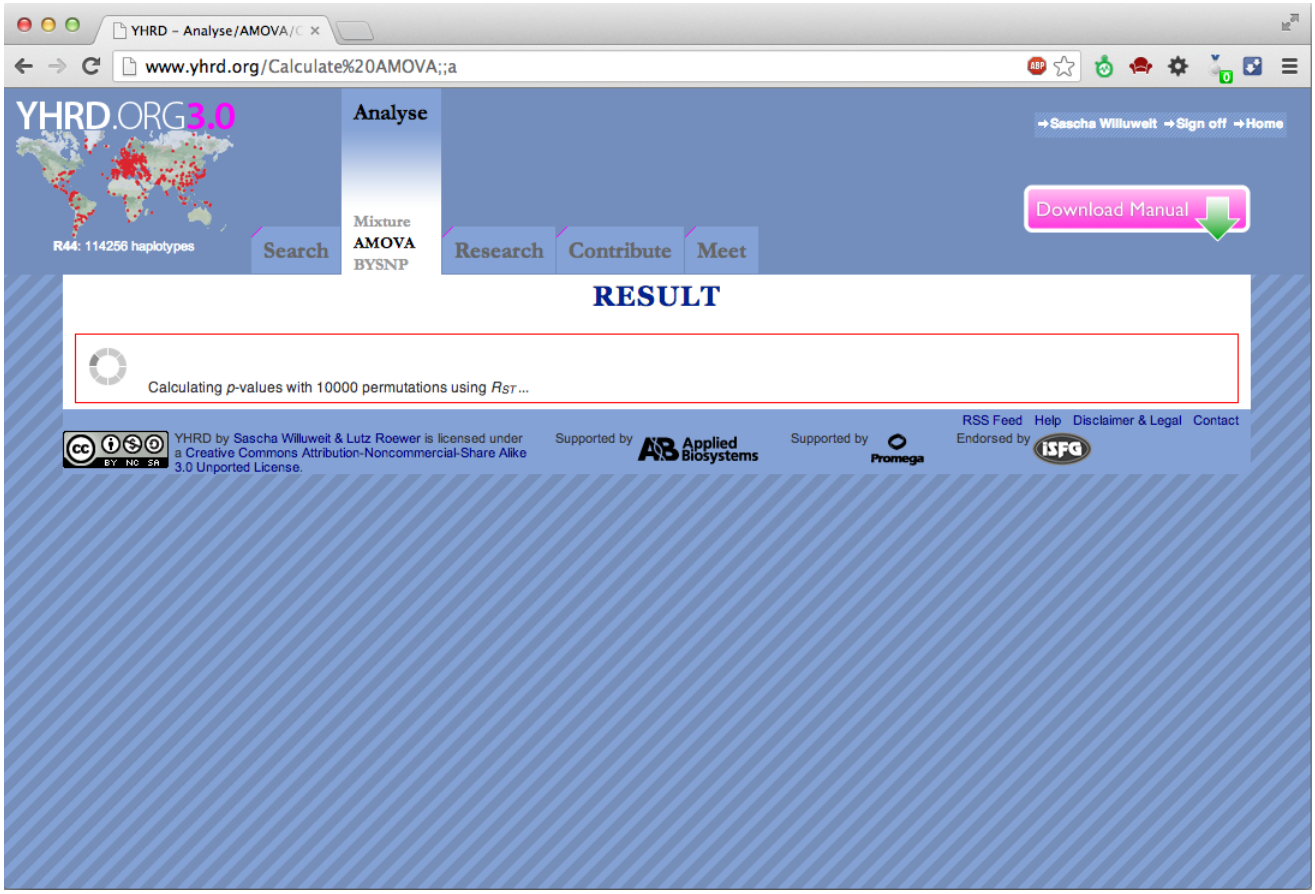


Figure 50: AMOVA: Result: Waiting...

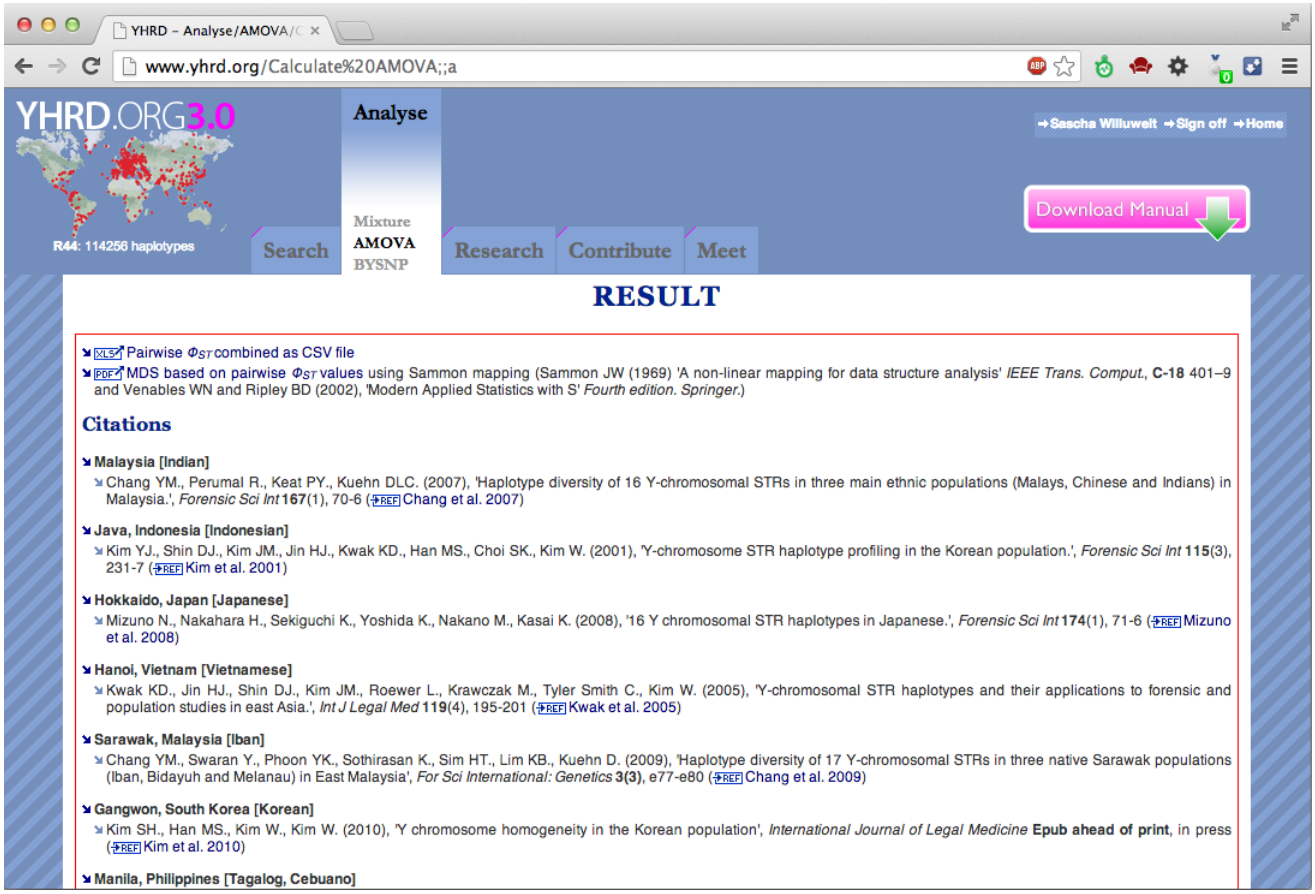
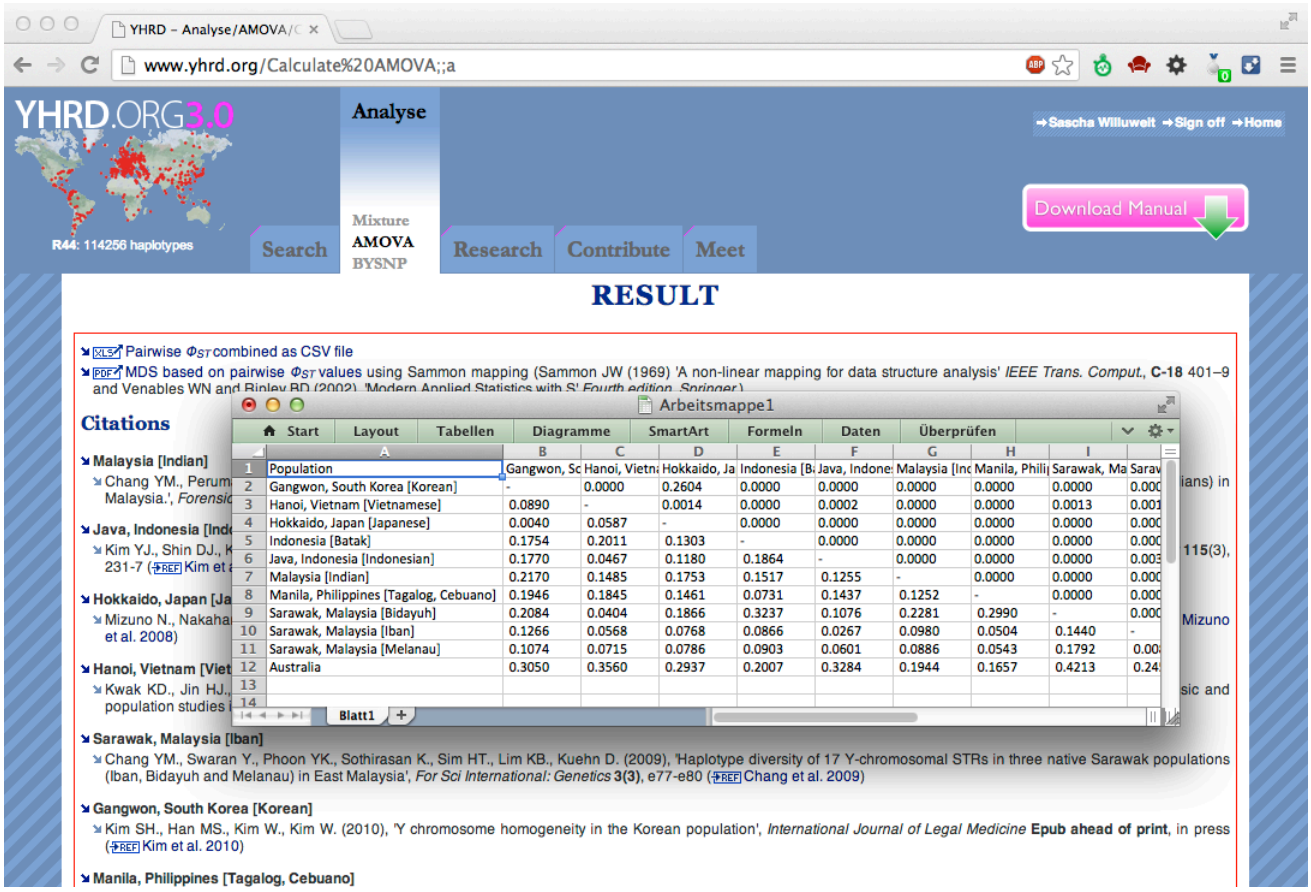


Figure 51: AMOVA: Result: Overview



	A	B	C	D	E	F	G	H	I
1	Population	Gangwon, Sc	Hanoi, Vietn	Hokkaido, Ja	Indonesia [B	Java, Indone	Malaysia [In	Manila, Phili	Sarawak, Ma
2	Gangwon, South Korea [Korean]	-	0.0000	0.2604	0.0000	0.0000	0.0000	0.0000	0.0000
3	Hanoi, Vietnam [Vietnamese]	0.0890	-	0.0014	0.0000	0.0002	0.0000	0.0000	0.0013
4	Hokkaido, Japan [Japanese]	0.0040	0.0587	-	0.0000	0.0000	0.0000	0.0000	0.0000
5	Indonesia [Batak]	0.1754	0.2011	0.1303	-	0.0000	0.0000	0.0000	0.0000
6	Java, Indonesia [Indonesian]	0.1770	0.0467	0.1180	0.1864	-	0.0000	0.0000	0.0000
7	Malaysia [Indian]	0.2170	0.1485	0.1753	0.1517	0.1255	-	0.0000	0.0000
8	Manila, Philippines [Tagalog, Cebuano]	0.1946	0.1845	0.1461	0.0731	0.1437	0.1252	-	0.0000
9	Sarawak, Malaysia [Bidayuh]	0.2084	0.0404	0.1866	0.3237	0.1076	0.2281	0.2990	-
10	Sarawak, Malaysia [Iban]	0.1266	0.0568	0.0768	0.0866	0.0267	0.0980	0.0504	0.1440
11	Sarawak, Malaysia [Melanau]	0.1074	0.0715	0.0786	0.0903	0.0601	0.0886	0.0543	0.1792
12	Australia	0.3050	0.3560	0.2937	0.2007	0.3284	0.1944	0.1657	0.4213
13									
14									

Figure 52: AMOVA: Result: Open Φ_{ST} table

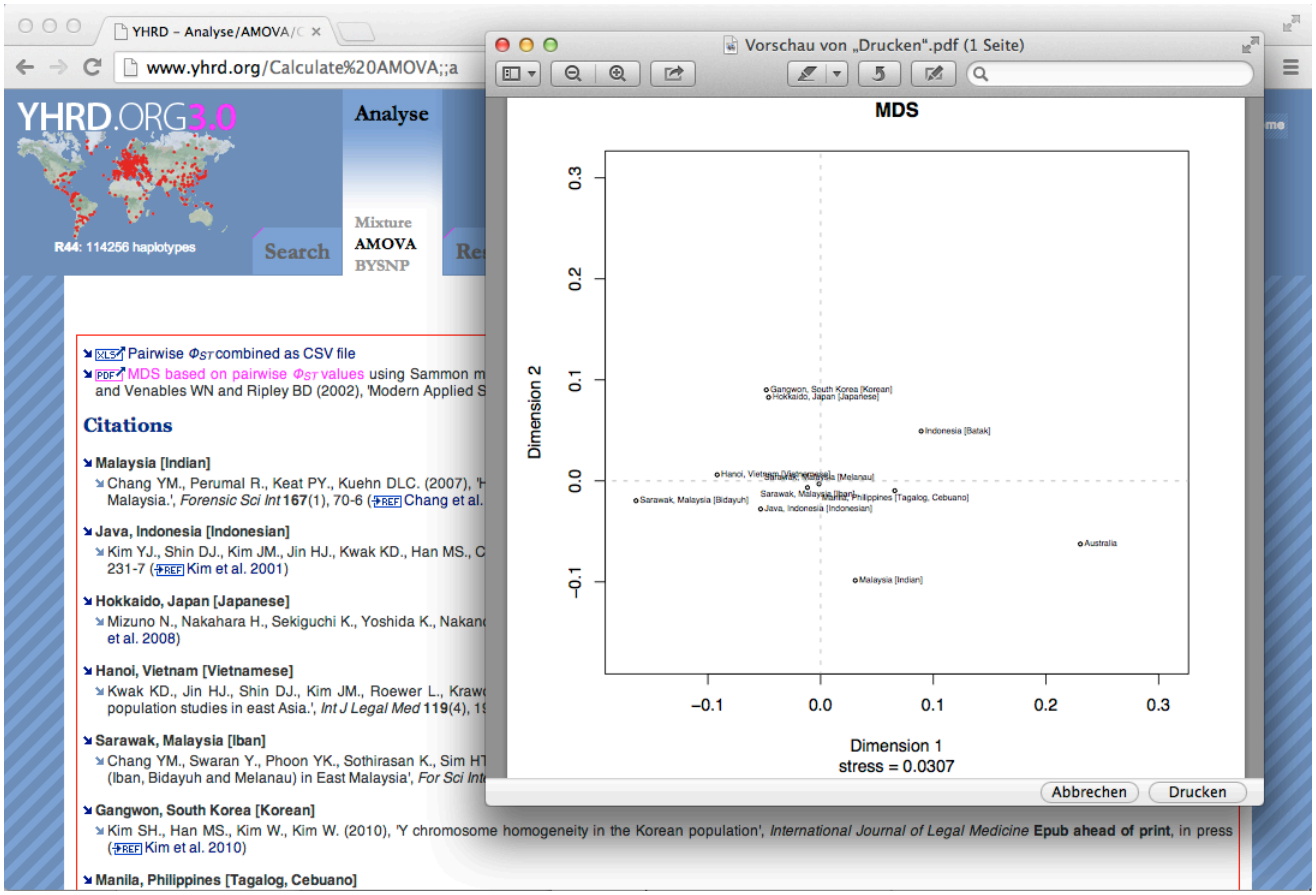


Figure 53: AMOVA: Result: Open MDS plot

3.2.2 Tool for mixture Interpretation – `http://www.yhrd.org/Analyse/Mixture`

This tool can be applied when a mixed trace (> 2 male contributors) and one person (suspect) which is included in the mixed profile should be analyzed.
We compare the trace (Y-STR data with alleles separated by semicolons) to the known putative contributor as well as to a number of unknown contributors.
There are the following limitations (also see Figure 54): All unknown persons should belong to the same ethnic group (Metapopulation) and they should be unrelated.
The result will be a likelihood of donor-ship vs. non-donor-ship of the suspect to the trace (Wolf et al. (2005), see Figures 55, 56 and 57).

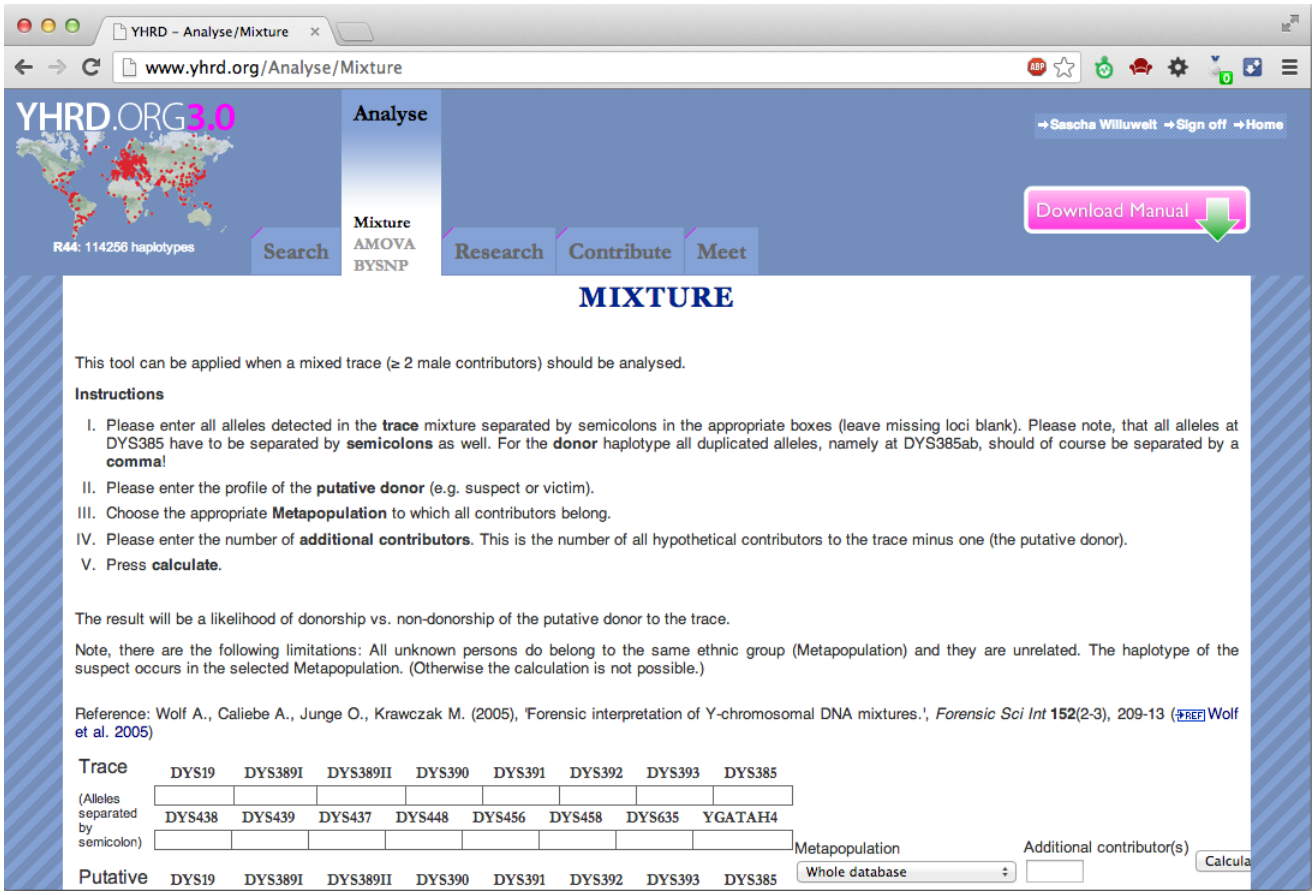


Figure 54: Mixture: Guide and list of requirements

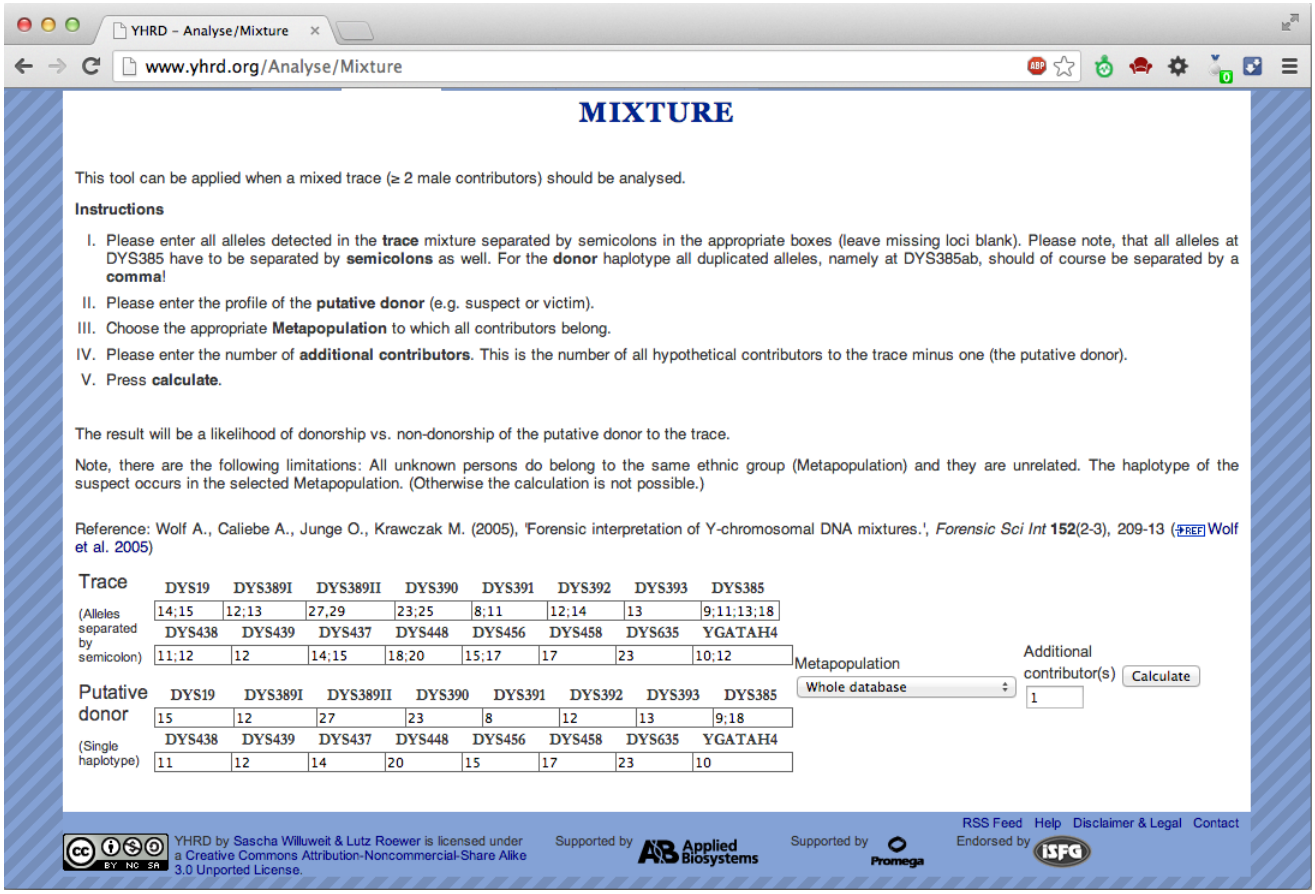


Figure 55: Mixture: Enter your data

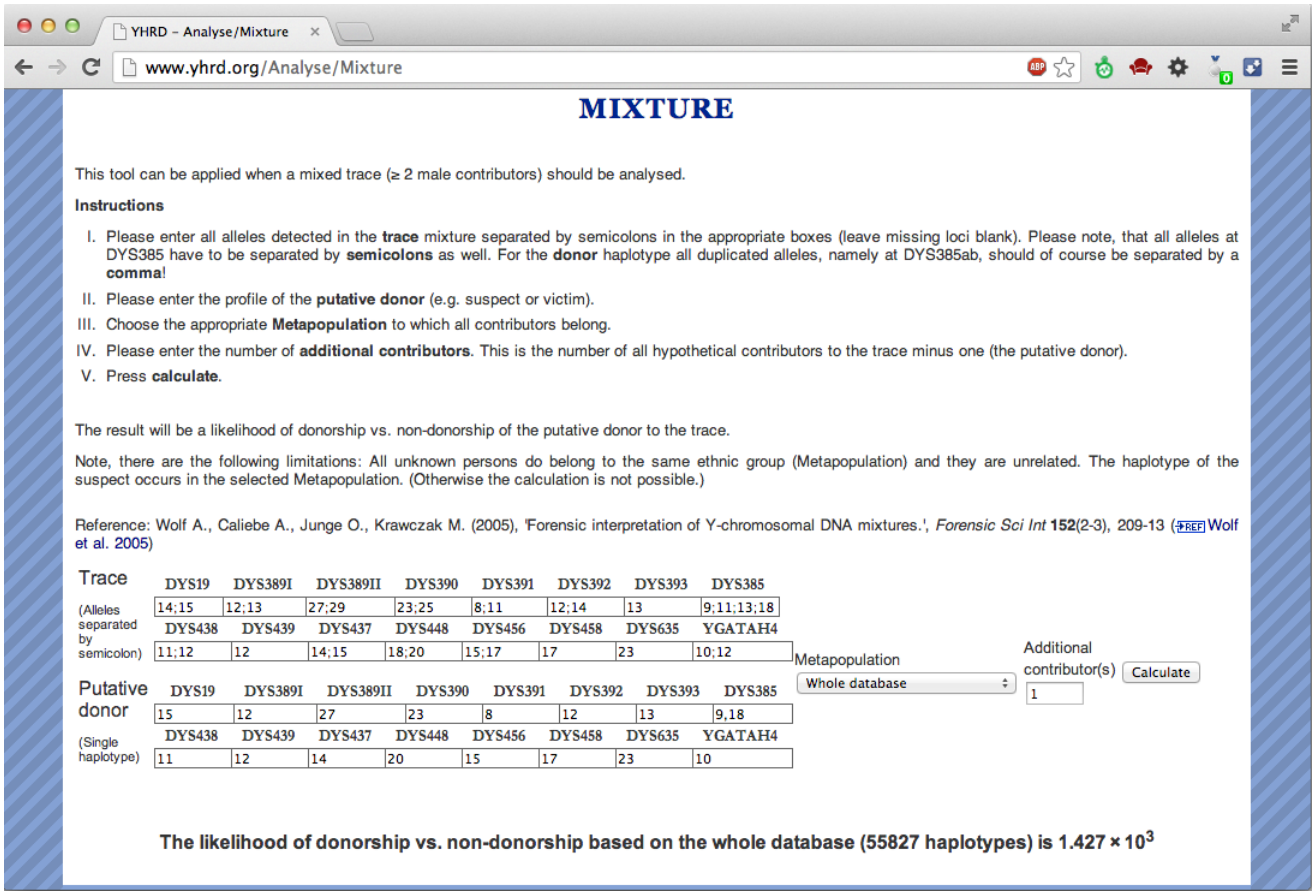


Figure 56: Mixture: Result I

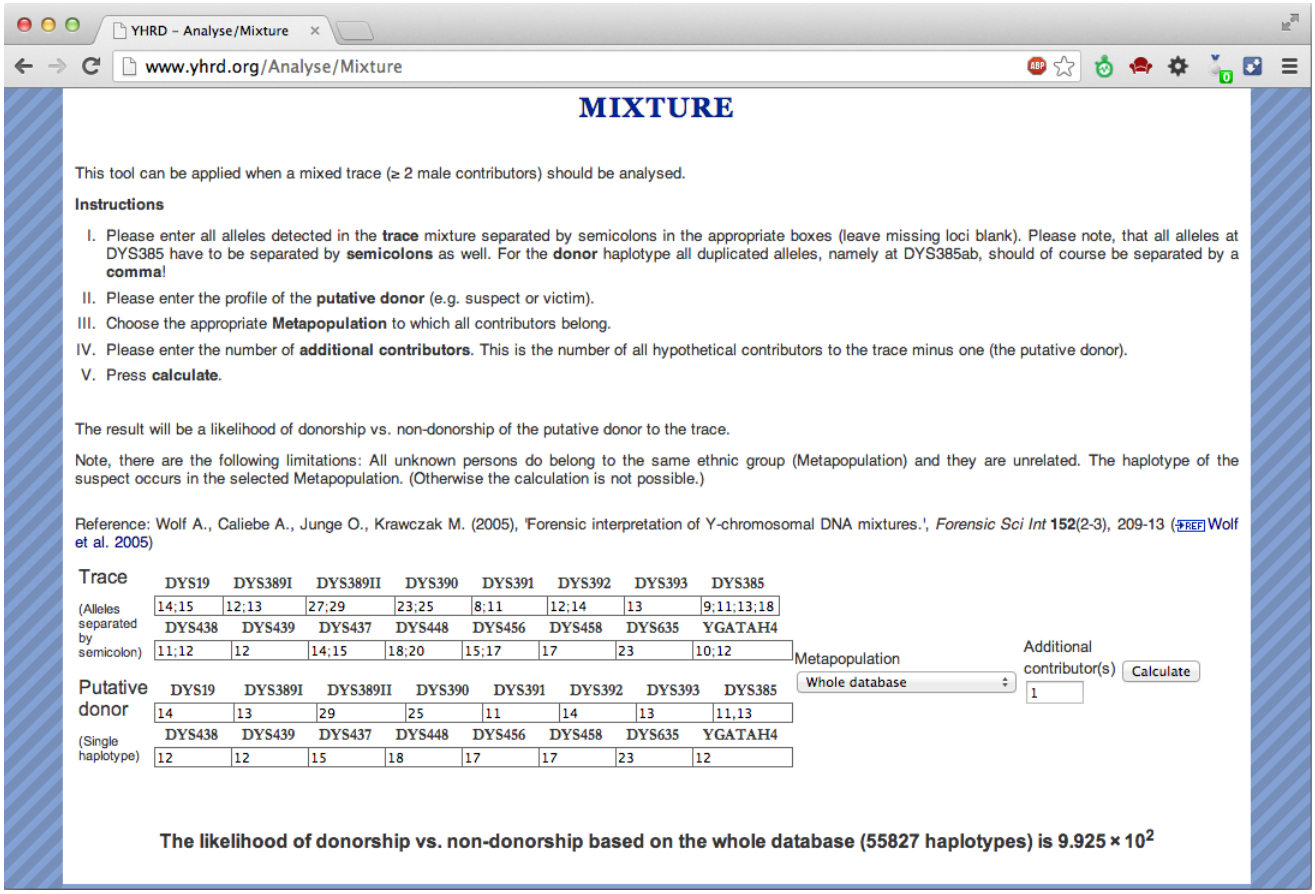


Figure 57: Mixture: Result II

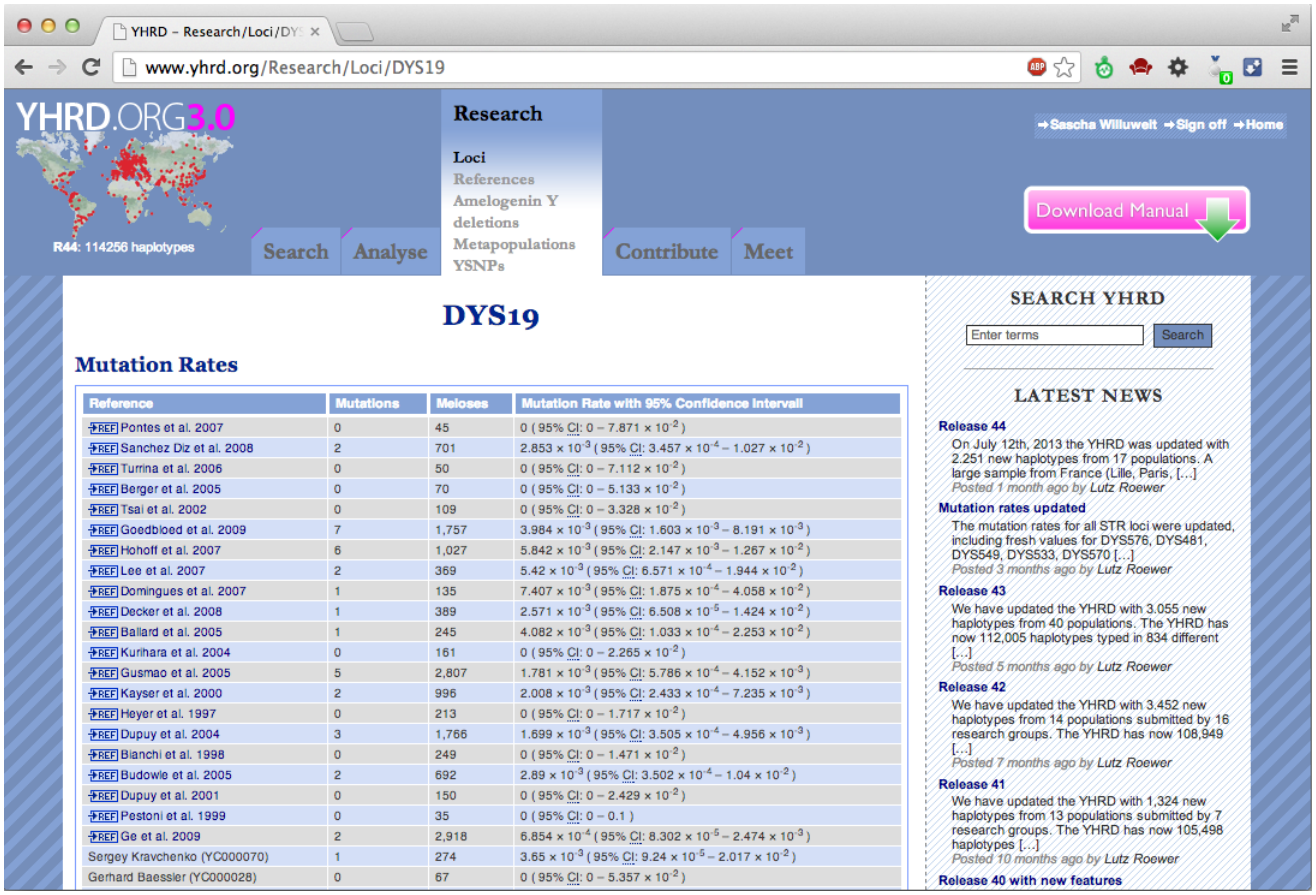


Figure 58: Locus DYS19 information: Mutation rates

3.3 Research – `http://www.yhrd.org/Research`

3.3.1 Loci – `http://www.yhrd.org/Research/Loci`

These pages provide a lot of information on genetic properties of those Y-STR markers and loci of which the forensic used Haplotypes are composed. In contrast to e.g. the database STRbase[9] all information (except the mutation rates which are collected from the available literature) is directly created from the database and is thus updated with each new release. A reference list is included in this section. Figures 58, 59, 60 and 61 show exemplary pages for locus DYS19.

[9] See `http://www.cstl.nist.gov/div831/strbase/`

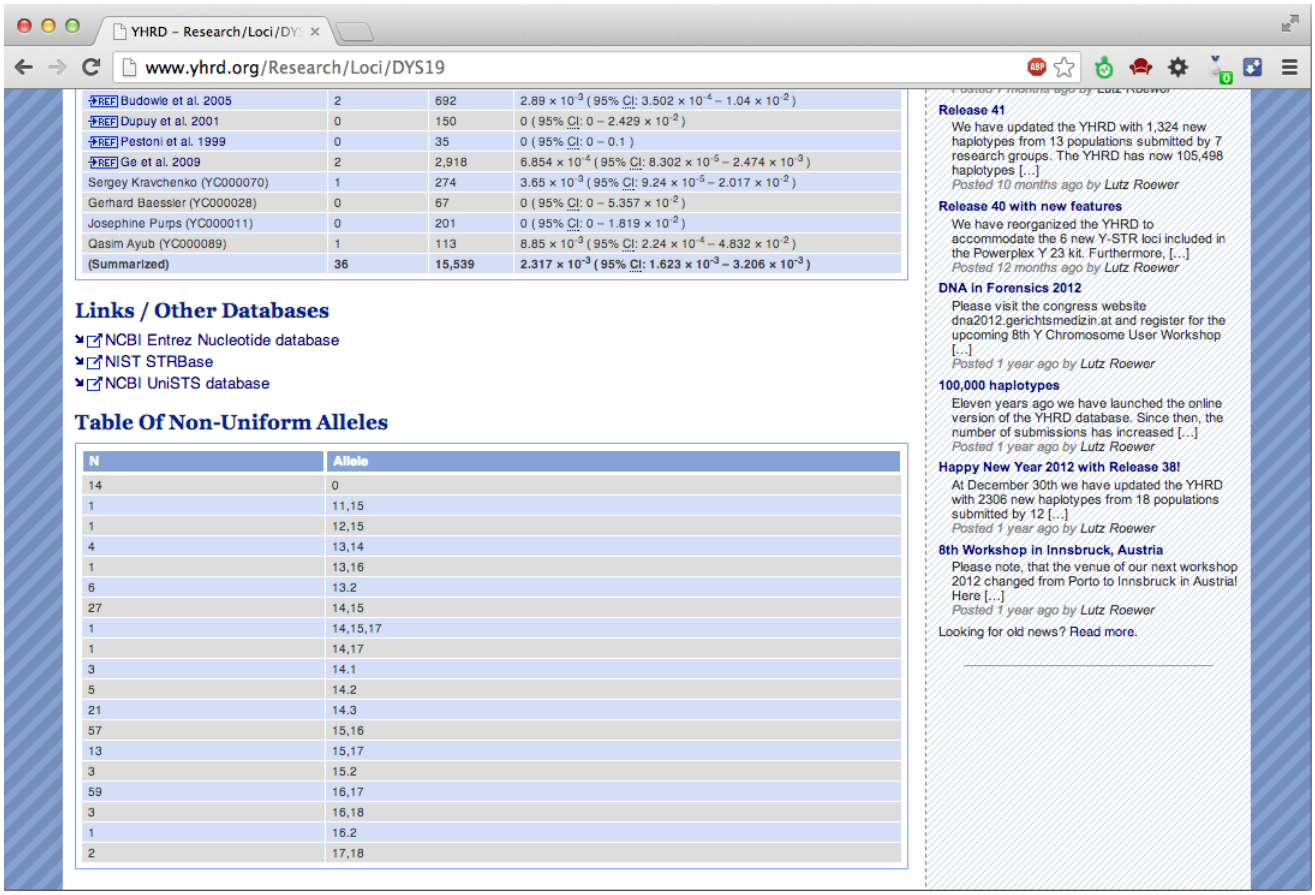


Figure 59: Locus DYS19 information: Links and table of non-uniform alleles

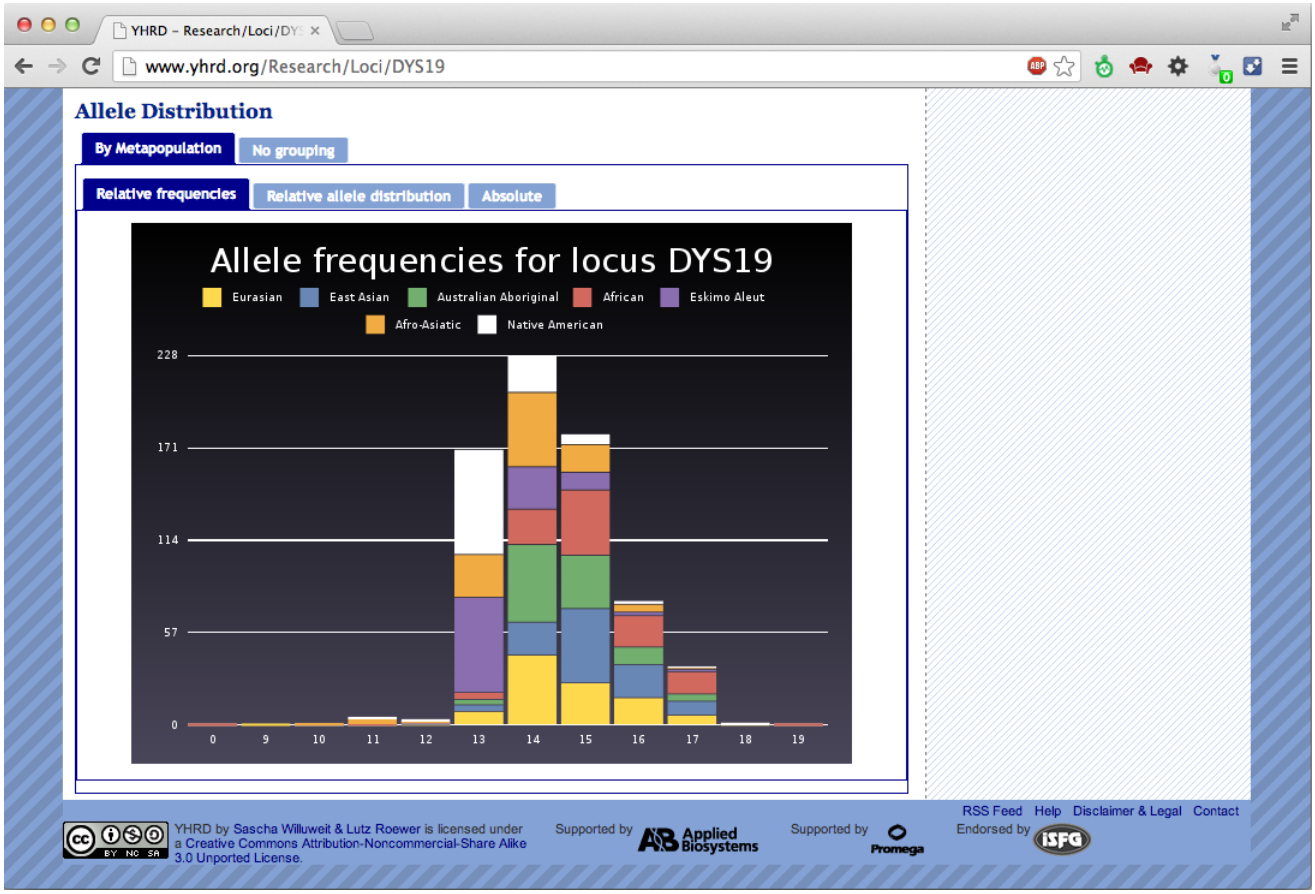


Figure 60: Locus DYS19 information: Relative distribution of alleles

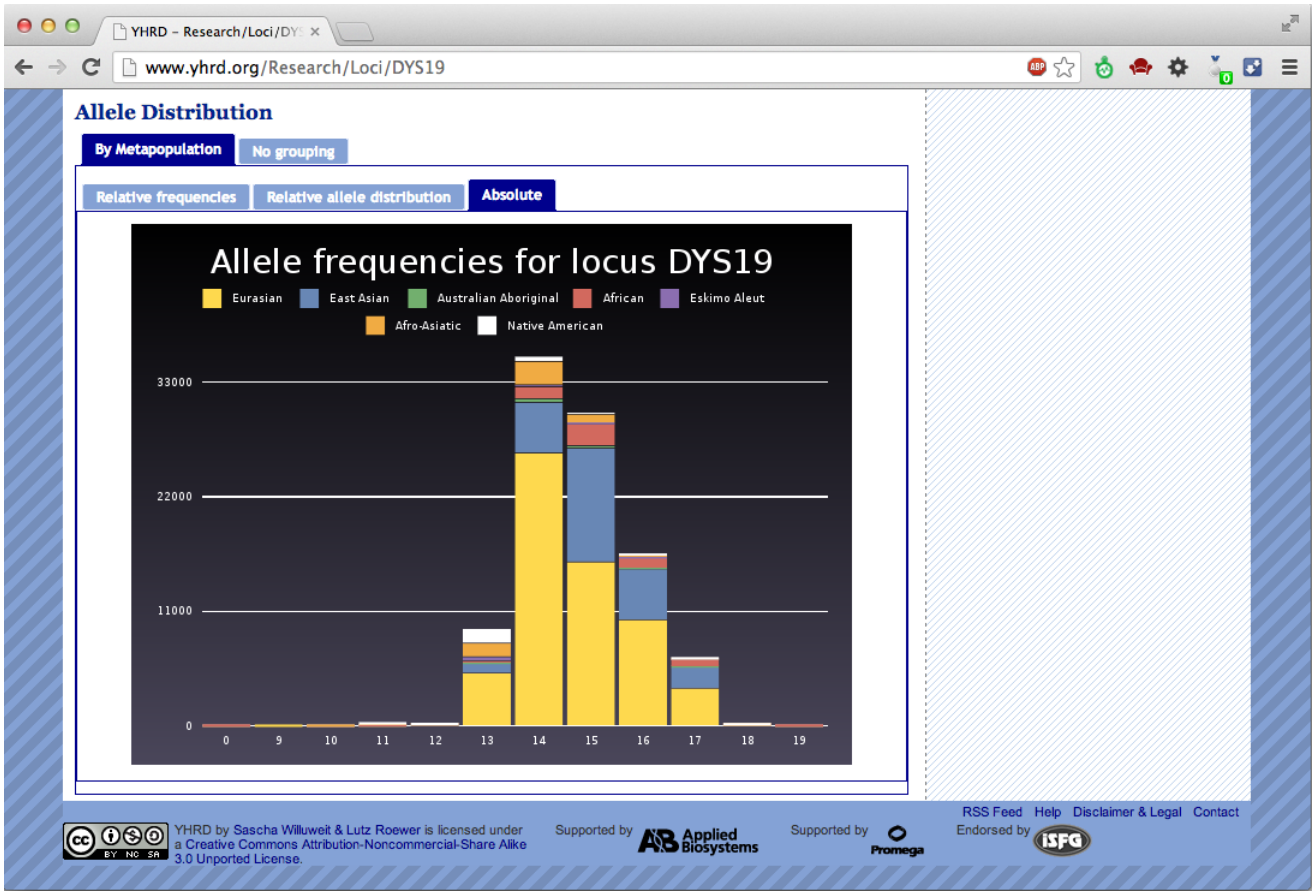


Figure 61: Locus DYS19 information: Absolute distribution of alleles

3.3.2 References – `http://www.yhrd.org/Research/References`

References are given in the "Research" pages of the website. Also projects using the YHRD database and links to other Y-STR databases are listed (see Figure 62).

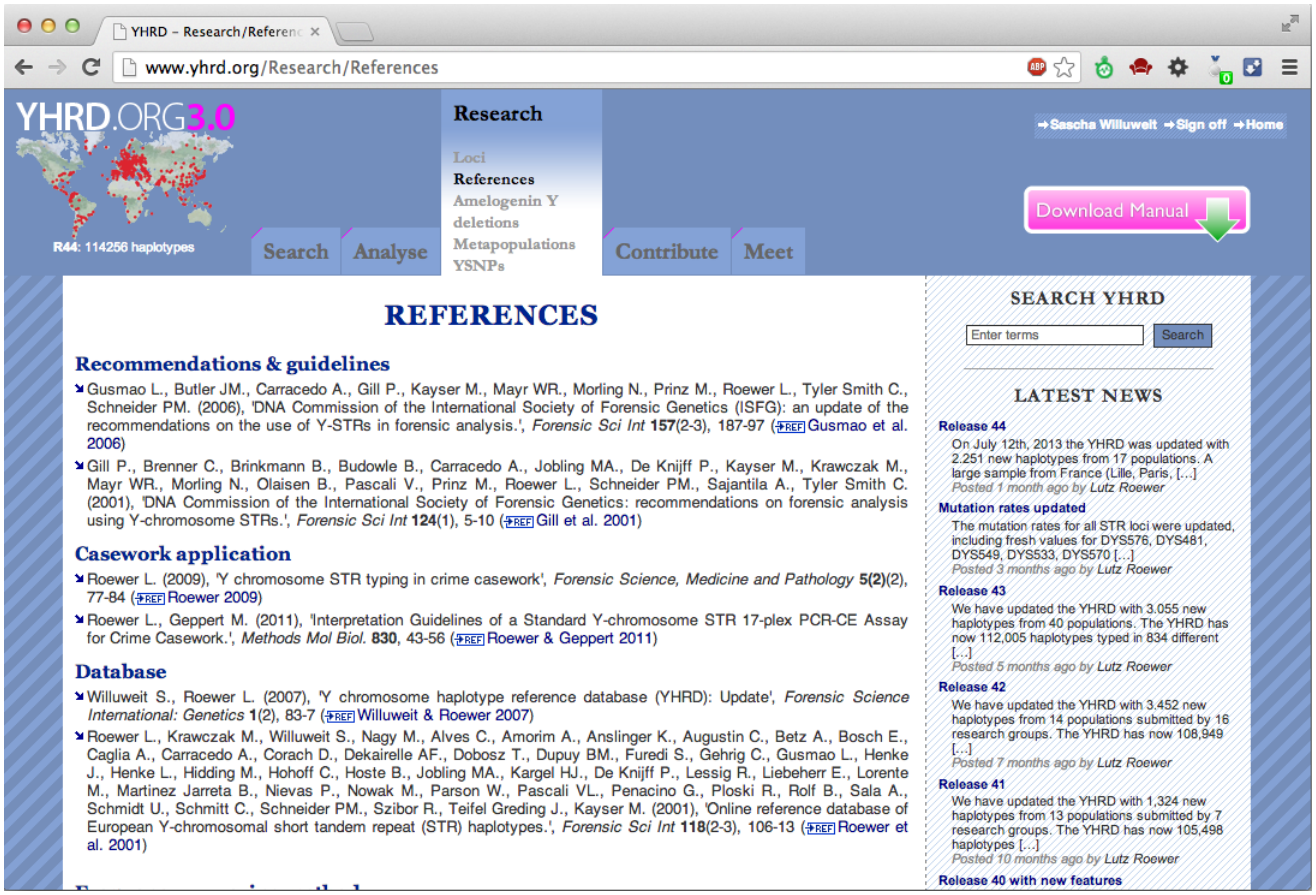
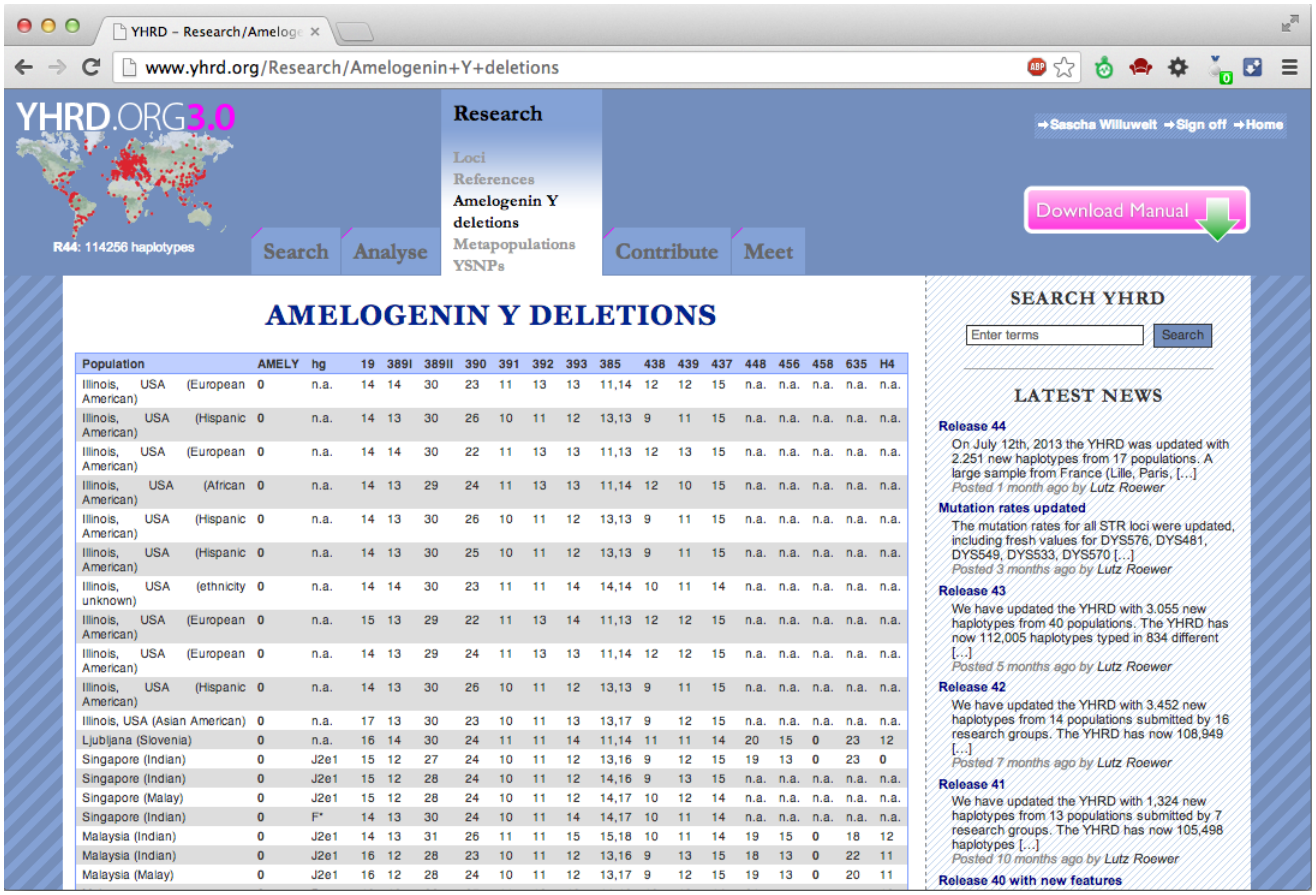


Figure 62: References summary page

3.3.3 **Amelogenin Y deletions** – `http://www.yhrd.org/Research/Amelogenin+Y+deletions`

Some contributors have submitted Amelogenin Y deleted chromosomes together with the haplotypes profil. These are listed under "Research" (see Figure 63).



Population	AMELY	hg	19	389I	389II	390	391	392	393	385	438	439	437	448	456	458	635	H4
Illinois, USA (European American)	0	n.a.	14	14	30	23	11	13	13	11,14	12	12	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (Hispanic American)	0	n.a.	14	13	30	26	10	11	12	13,13	9	11	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (European American)	0	n.a.	14	14	30	22	11	13	13	11,13	12	13	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (African American)	0	n.a.	14	13	29	24	11	13	13	11,14	12	10	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (Hispanic American)	0	n.a.	14	13	30	26	10	11	12	13,13	9	11	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (Hispanic American)	0	n.a.	14	13	30	25	10	11	12	13,13	9	11	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (ethnicity unknown)	0	n.a.	14	14	30	23	11	11	14	14,14	10	11	14	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (European American)	0	n.a.	15	13	29	22	11	13	14	11,13	12	12	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (European American)	0	n.a.	14	13	29	24	11	13	13	11,14	12	12	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (Hispanic American)	0	n.a.	14	13	30	26	10	11	12	13,13	9	11	15	n.a.	n.a.	n.a.	n.a.	n.a.
Illinois, USA (Asian American)	0	n.a.	17	13	30	23	10	11	13	13,17	9	12	15	n.a.	n.a.	n.a.	n.a.	n.a.
Ljubljana (Slovenia)	0	n.a.	16	14	30	24	11	11	14	11,14	11	11	14	20	15	0	23	12
Singapore (Indian)	0	J2e1	15	12	27	24	10	11	12	13,16	9	12	15	19	13	0	23	0
Singapore (Indian)	0	J2e1	15	12	28	24	10	11	12	14,16	9	13	15	n.a.	n.a.	n.a.	n.a.	n.a.
Singapore (Malay)	0	J2e1	15	12	28	24	10	11	12	14,17	10	12	14	n.a.	n.a.	n.a.	n.a.	n.a.
Singapore (Indian)	0	F*	14	13	30	24	10	11	14	14,17	10	11	14	n.a.	n.a.	n.a.	n.a.	n.a.
Malaysia (Indian)	0	J2e1	14	13	31	26	11	11	15	15,18	10	11	14	19	15	0	18	12
Malaysia (Indian)	0	J2e1	16	12	28	23	10	11	12	13,16	9	13	15	18	13	0	22	11
Malaysia (Malay)	0	J2e1	16	12	28	24	10	11	12	13,17	9	12	15	19	13	0	20	11

Figure 63: Amelogenin Y deletions

3.3.4 **Metapopulations** – `http://www.yhrd.org/Research/Metapopulations`

In this section all predefined Metapopulations (see Figure 64) are on display with maps describing the catchment area, the geographic dispersal (see Figures 65 and 66) and a list of assigned population samples (see Figure 67).

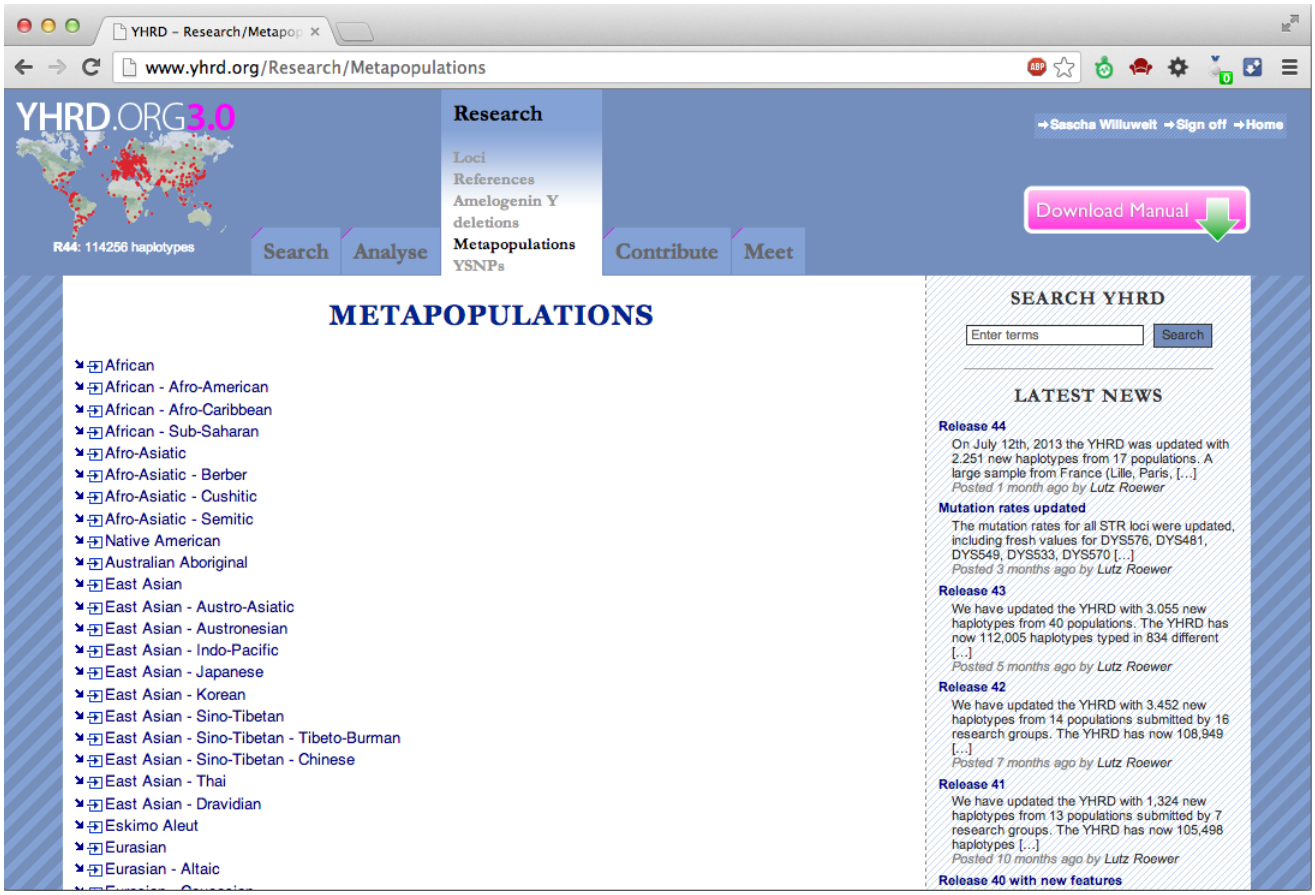


Figure 64: Predefined Metapopulations

Figure 65: Catchment area of the Native American Metapopulation

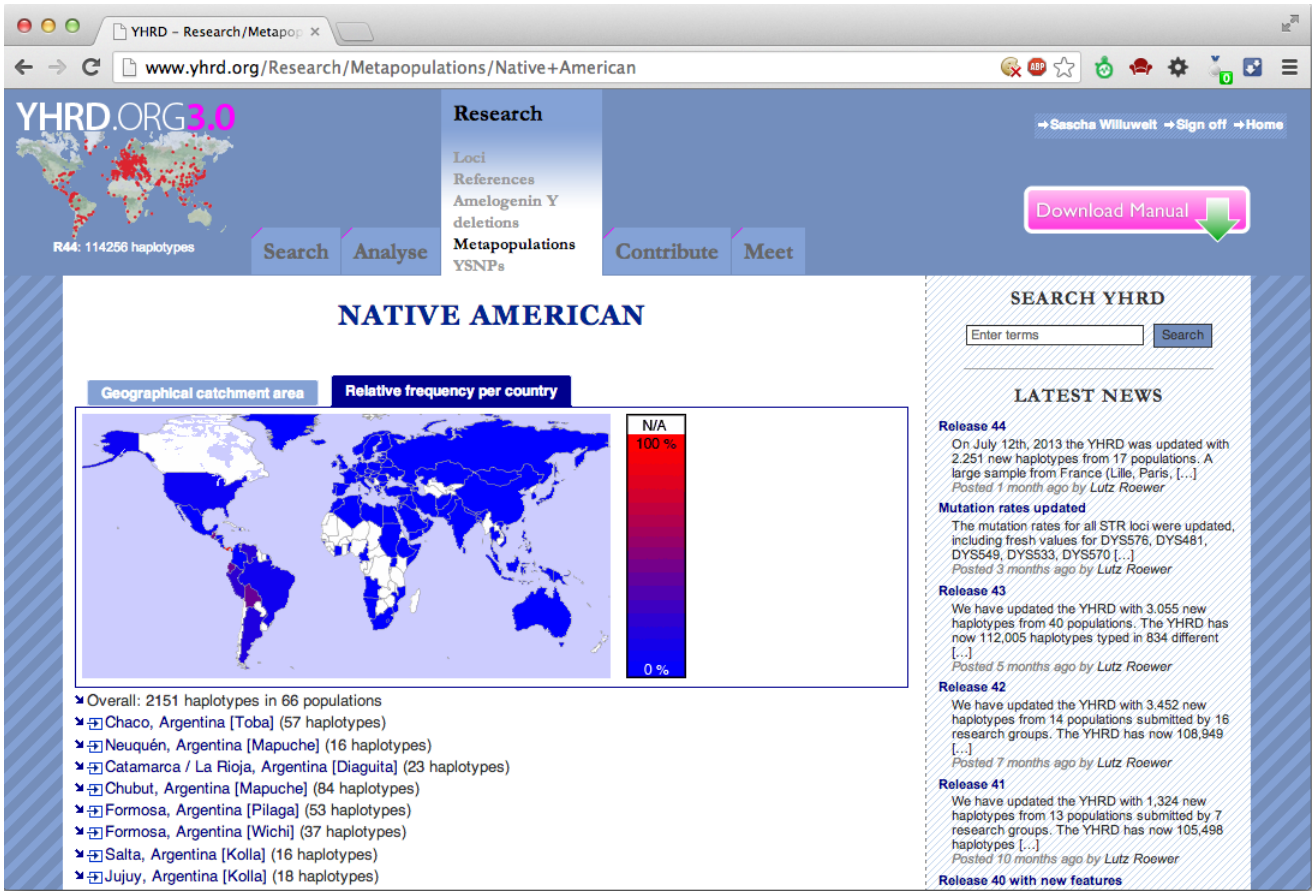


Figure 66: Geographic distribution of the Native American Metapopulation

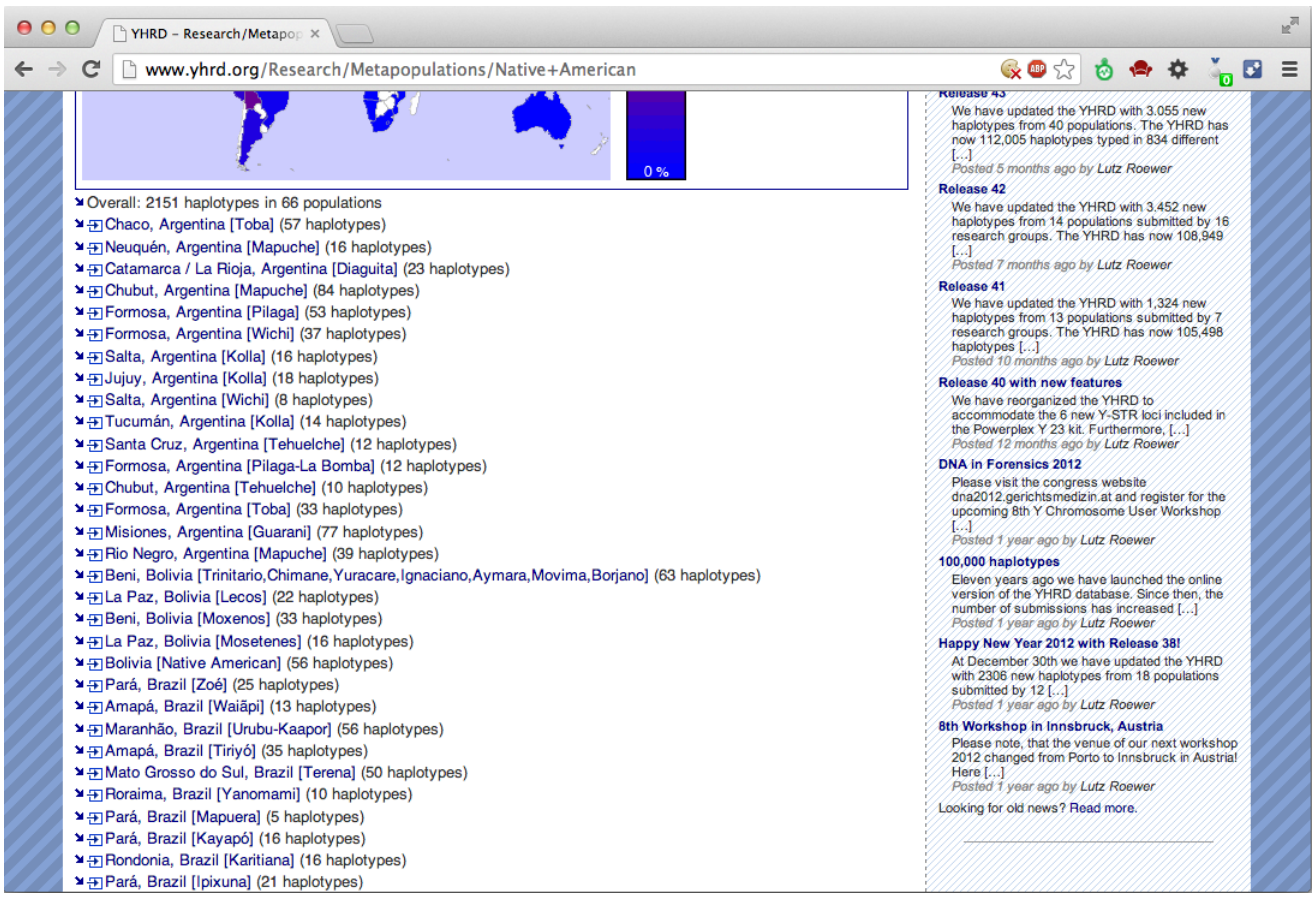


Figure 67: List of populations assigned to the Native American Metapopulation

Figure 68: Y-SNP Fact Sheet

3.3.5 YSNPs – `http://www.yhrd.org/Research/YSNPs`

The Y-SNP section provides all molecular an phylogenetic information on Y-SNP markers and the assigned Haplogroup branches (see Figure 68).

Haplogroup branch section The first part of this section indicates the placement of the haplogroup branch within the hierarchy of the phylogenetic tree. In addition, the defining markers, the curator and timestamps of creation and modification are displayed. Finally, you get a map of the global distribution of the Haplogroup. The dispersal is depicted country-wise, graded from blue (not observed) to red (all individuals carry this Haplogroup) (see Figure 69).

Haplogroup marker section The content may vary, but in most instances the following information is displayed: Marker name aliases (if any), assigned Haplogroup branch, specific type of mutation, flanking sequence with links to BLAST (partial or complete sequence) against the NCBI human genome, NCBI dbSNP link to SNP accession number, references, list of commercial services offering to analyse this marker, curator, timestamps and lab protocol(s) (see Figures 70 and 71).

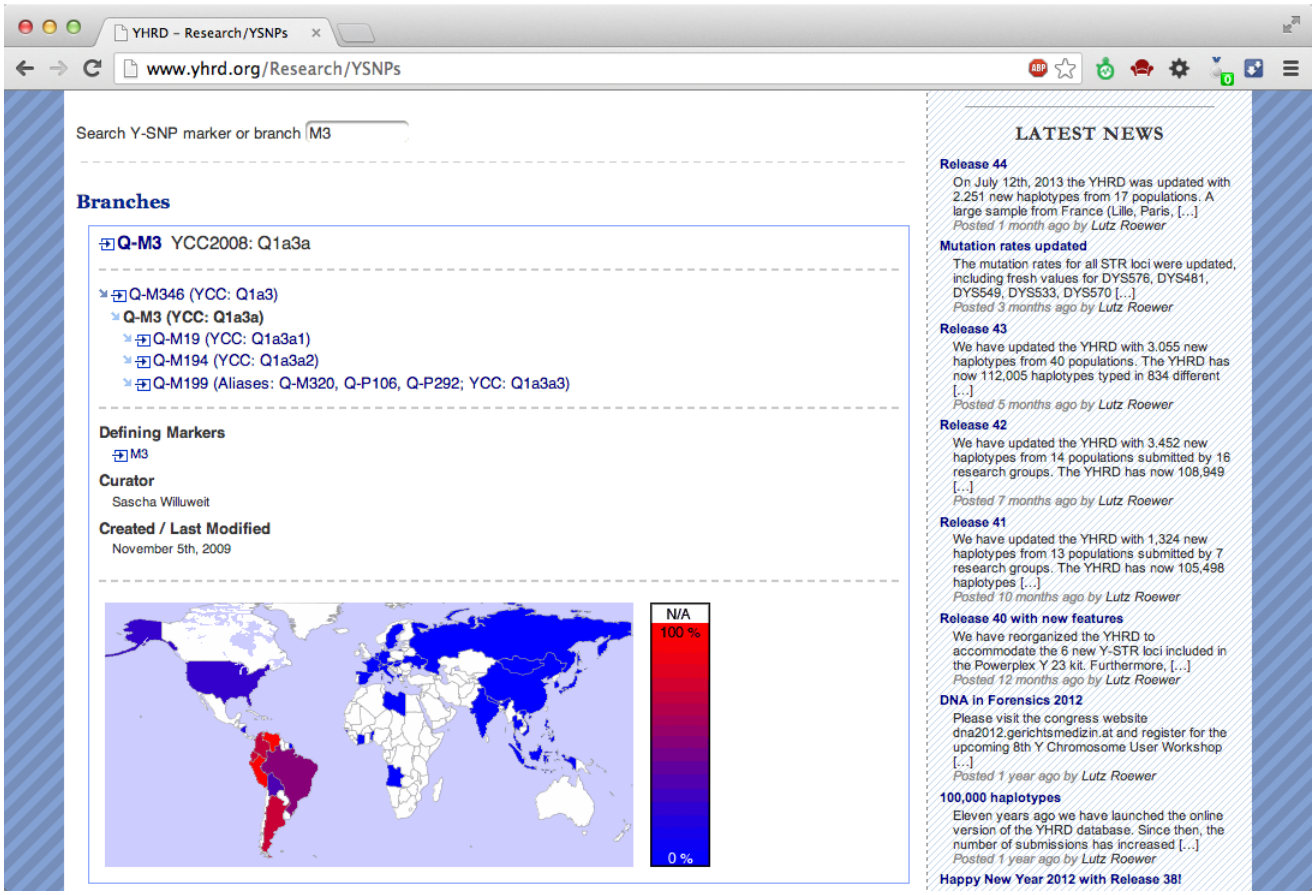


Figure 69: Y-SNP Fact Sheet: Branch section

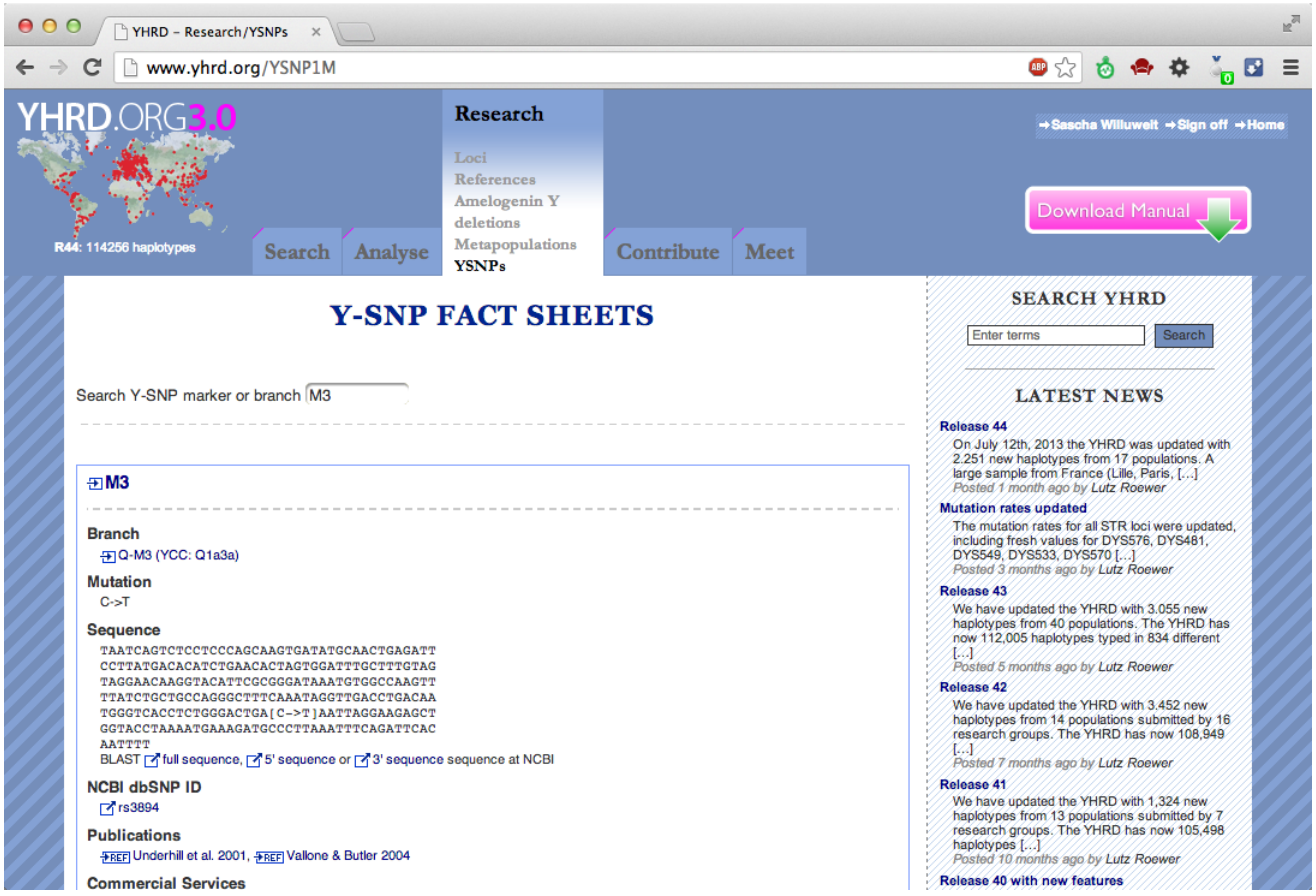


Figure 70: Y-SNP Fact Sheet: Marker section

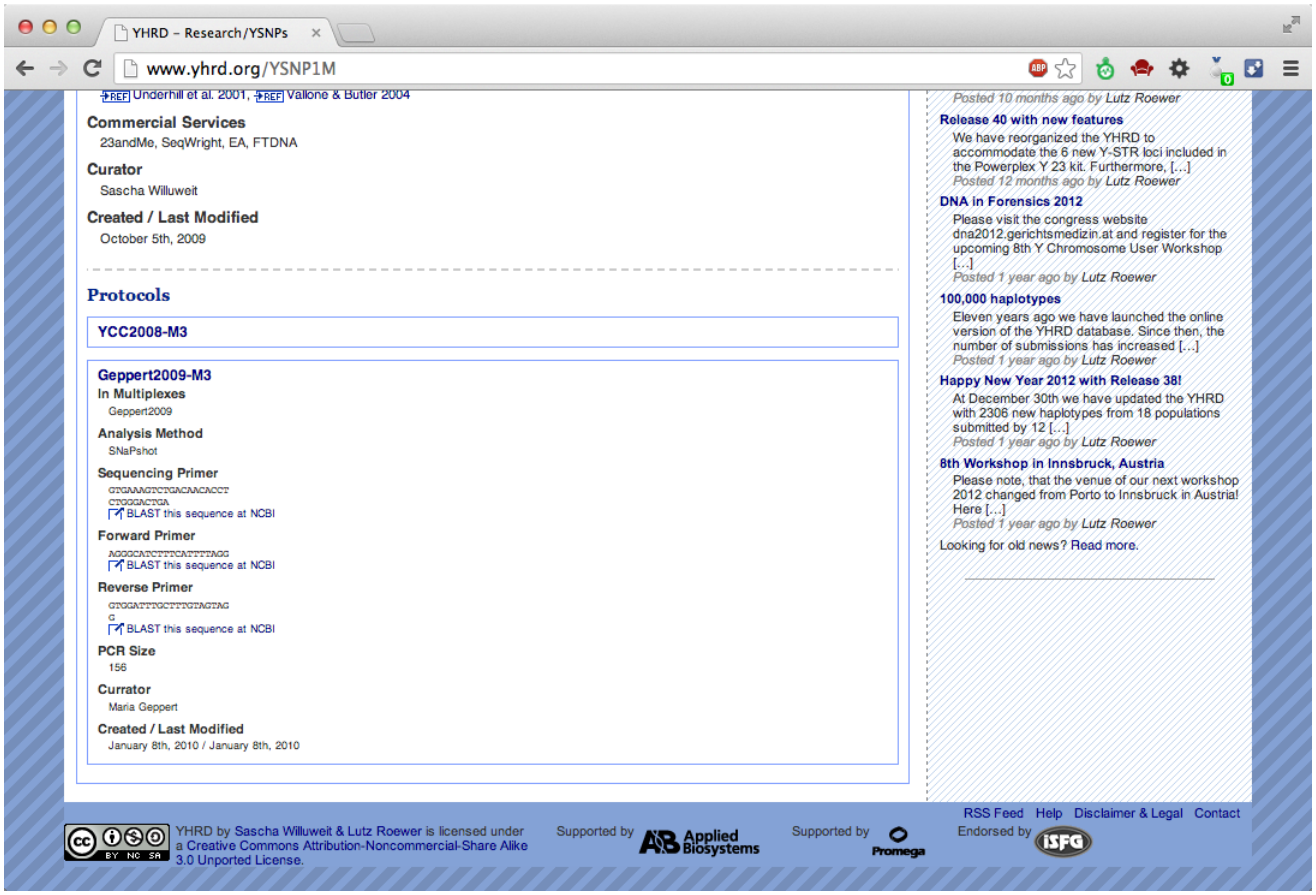


Figure 71: Y-SNP Fact Sheet: Marker section, protocol details

3.4 Contribute – `http://www.yhrd.org/Contribute`

This page explains in detail the submission of population samples to the YHRD and the procedure to receive an accession number (mandatory for publication in FSI:Genetics[10]) (see Figure 72).

[10] See `http://www.fsigenetics.com`

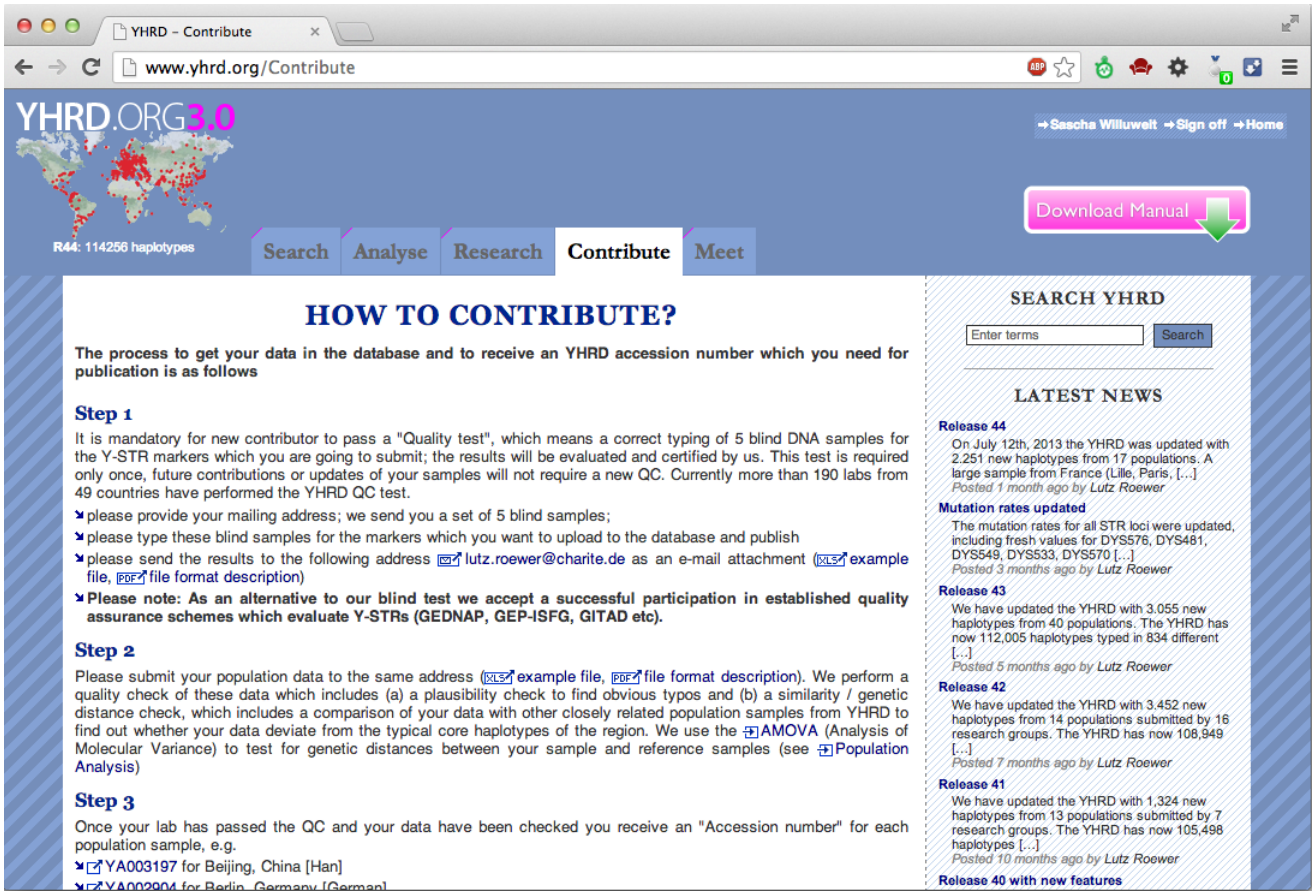


Figure 72: Contribution guideline

3.5 Authorization

YHRD only allows two searches and none AMOVA calculation without registration (see Figure 74). After registration[11] and login[12] (see Figure 73) you have access to the database (20 searches per day), could perform AMOVA (2 per day) and use all tools. After 30 min of inactivity your session expires and you need to login again. Within your personal profile page (by clicking on your name on the upper left of the site like in Figure 75) you will find all your searches and are able to change your name or email address.

[11] See `http://www.yhrd.org/Register`
[12] See `http://www.yhrd.org/Login`

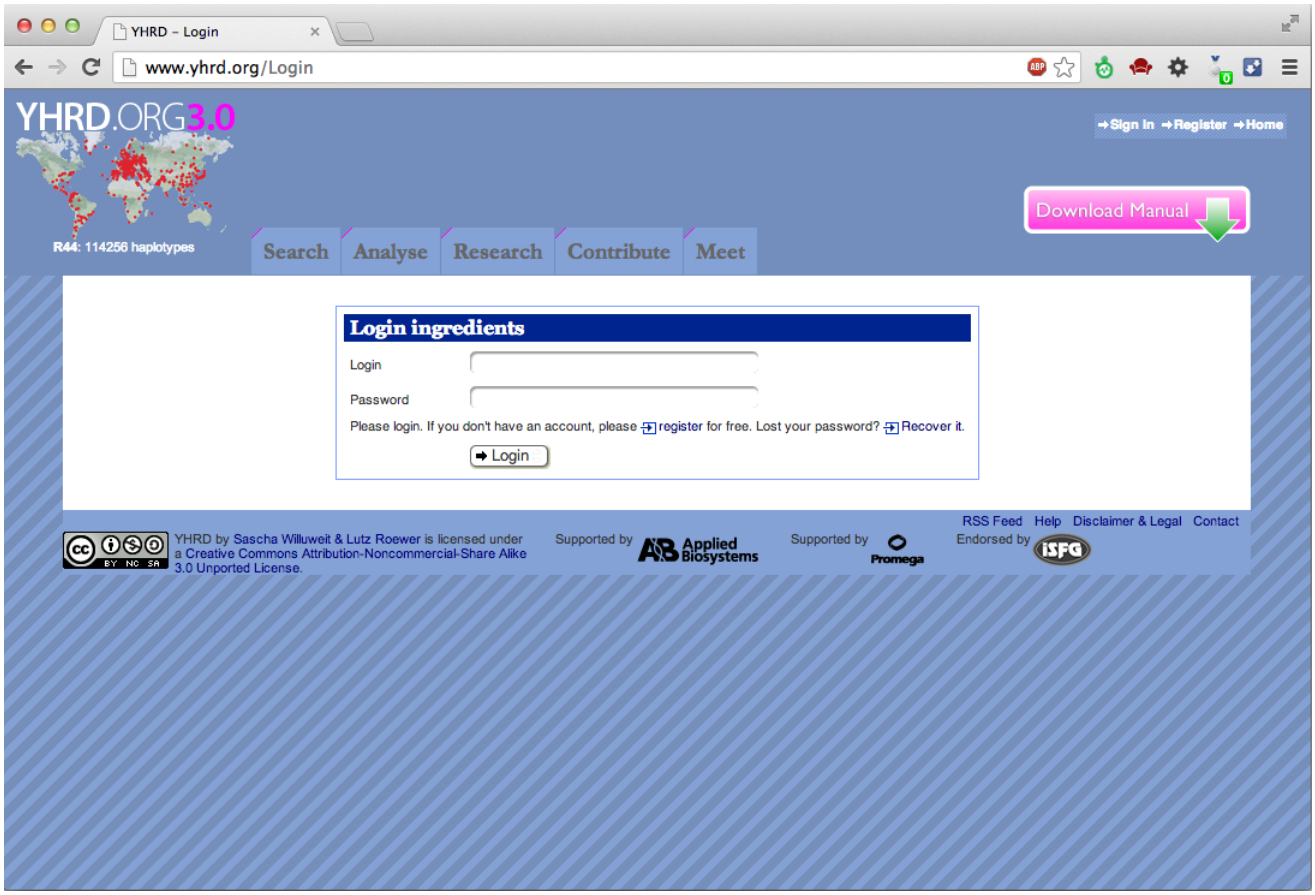


Figure 73: Sign in

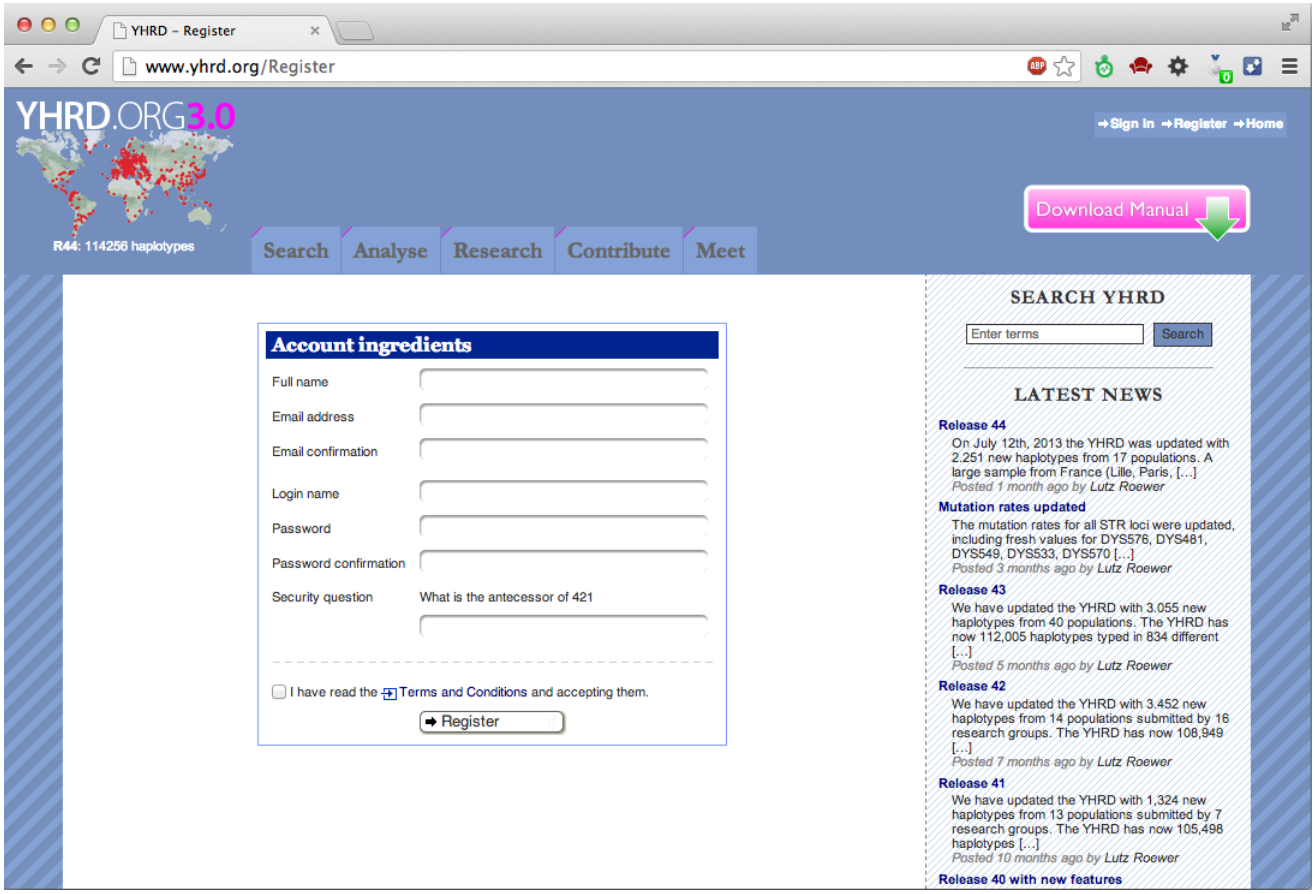


Figure 74: Register for an account

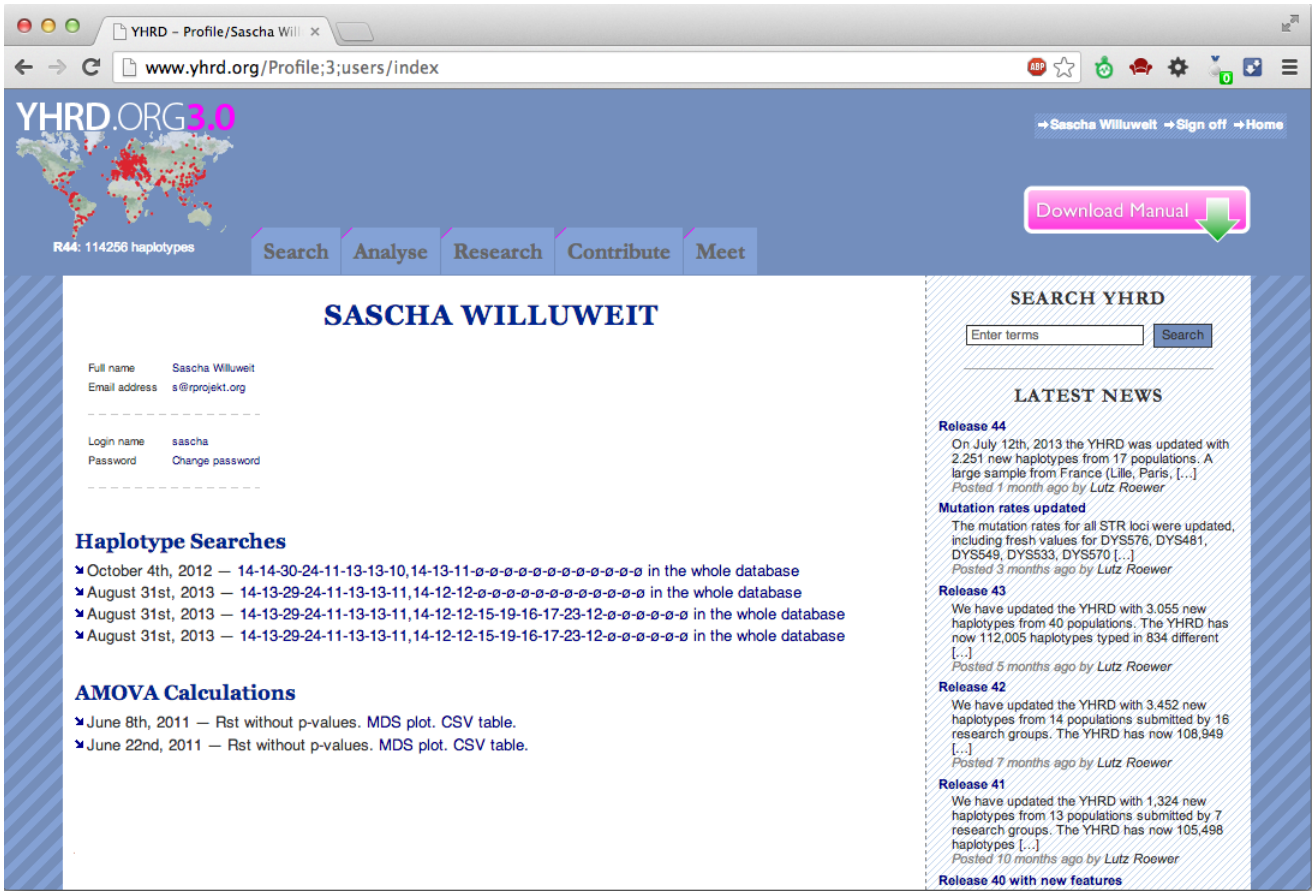


Figure 75: Users profile page

4 Glossary

4.1 Metapopulations

In population genetics the term "Metapopulation" describes discrete spatially distributed population groups which are interconnected by geneflow and migration (Hanski and Gilpin (1997)).
By analogy, the term Metapopulation is used in forensic genetics to describe a set of geographically dispersed populations which are connected by geneflow and are thus more similar within the Metapopulation (MP) than to groups outside the MP (Willuweit et al. (2007)).
Match calculations performed on basis of such pooled population data should not be significantly affected by sub-population structure. However, large genetic distances between Y chromosome Metapopulations affect match calculations with respect to the used Metapopulation.
Sampled individuals can be grouped as a population according to several shared characteristics such as nationality, geography, language affiliation or ethnic ancestry. Notoriously, these definitions often cannot conclusively define individuals ancestry to a group (i.e. a population). Therefore none of them should be used alone or regarded as superior to other shared proxies.
A new pooling approach in forensic genetics using Y chromosome STR databases is to cluster individual samples to groups according to their phylogenetic ancestry (e.g. Y-SNP defined Haplogroups).
Currently the YHRD database recognizes four separate Metapopulation structures; national (see Section 4.1.1), continental (see Section 4.1.2), linguistic/ethnic (see Section 4.1.3) and phylogenetic affiliation with several categories within (see Section 4.1.4).

4.1.1 National

The concept of pooling data to build "national databases" has a very straightforward explanation: law enforcement agencies and forensic services rely on their national population to built reference databases. In most instances offenders and victims stem from the national population, and their genetic profiles should thus be represented in the database. In countries like USA, Brazil, UK or China which are characterized by strong population substructure national reference databases are often built on basis of a historical concept of ethnic affiliation, e.g. the US population is sub-structured in a Caucasian, African, Hispanic, Asian and Native American populations or UK differentiates English, Afro-Caribbean, Indo-Pakistani and Chinese. National databases due to their importance

in national legislation are thus searchable in the YHRD. Each national Metapopulation in the YHRD comprises all individuals sampled in a particular country regardless of the ancestry of the individuals.

4.1.2 Continental

Continental Metapopulations in the YHRD comprises all individuals sampled in a particular continent regardless of their ancestries. The YHRD defines six continental Metapopulations following the United Nations classification of geographical regions: Africa, Asia, Europe, Latin America, North America, Oceania.

4.1.3 Linguistic/Ethnic

The Metapopulation structure built on basis of "ethnicity/linguistic affiliation" takes to a larger extent the ancestry of sampled individuals into account. "Ancestry" is an term collating historical, cultural, geographical and linguistic categories. Of course, a Metapopulation concept on basis of "ethnicity" is by no means ideal, fully rational or fully translatable, but simply takes the fact into account that on a global level categories other than "nation" or "geography" far better describe the observed genetic clustering and inhomogeneity of Y chromosome patterns.

For a global reference database the "major language group" criterion seems most appropriate to group data by taking the ancestry into account and produce subdatabases with respect to genetic similarity. The reasoning in doing so is twofold: first, language is often an inherited cultural trait and thus the language phylae show strong similarity to genetic traits including the Y chromosome. Second, since languages are well examined by science and mostly understood by the public due to the long tradition of language research, the linguistic terminology is in principal more understandable and translatable into practice than their genetic pendant. Aside from the pure linguistic categorization (e.g. the Altaic language family comprising people speaking Turk and Mongol languages) we took also unifying geographic criterions (Sub-Saharan Africa comprising speakers of different African language groups which live south of the Sahara).

It is important to state, that the current "Metapopulation" structure is an a-priori categorization which needs a continuous evaluation and verification by means of statistical methods to quantify the genetic similarity/dissimilarity between the samples. While the current categorization of eight large Metapopulations gains some support from genetic distance analysis done on basis of 41,000 Haplotypes (Willuweit et al. (2007), see Figure 77) a further subdivision of the "Eurasian - European Metapopulation" was implemented solely on basis of Y-STR Haplotypes. The analysis of 12,000 European Haplotypes by AMOVA demonstrates that three largely homogeneous pools of European Haplotypes exist: the Western, the Eastern and the Southeastern Metapopulation (Roewer et al. (2005)). A search option for these subgroups is implemented in the YHRD. Further genetic distance analysis will show, which groups can be abandoned, further divided or need a new definition.

Currently the YHRD has eight non-overlapping Metapopulations (see Table 2): Admixed, African (see Figure 78), Afro-Asiatic (see Figure 80), Amerindian, Australian Aboriginal, East Asian (see Figure 76), Eskimo-Aleut, and Eurasian (see Figure 77). Some of these Metapopulations are further subdivided, e.g. Eurasian into six subcategories, from which European subgroup splits further into three groups of Western, Eastern and Southeastern Europeans.

Metapopulation	Haplotypes	Definition
All	114256	The assembly of all metapopulations builds the YHRD database. Language, geography and Y chromosome markers are used to define metapopulations.
African	5223 (300)	This MP comprises populations of recent Sub-Saharan ancestry, either living in the Subsaharan part of Africa or migrating to other continents (e.g. the US or the Caribbean)
Afro-American	2784	
Afro-Caribbean	489	
Sub-Saharan	1650	
Afro-Asiatic	4949 (0)	This MP comprises populations from North Africa and the Arabian countries of the Near East. The name "Afro-Asiatic" describes the language group where all these populations belong to.
Berber	354	
Cushitic	201	
Semitic	4394	
Australian Aboriginal	766	
East Asian	10384 (0)	This MP comprises all populations from the eastern part of Asia which are not Eurasian. The main language groups subsumed under "East Asian" are Austro-Asiatic (e.g. Vietnamese), Austronesian (e.g. Malay), Thai, Sino-Tibetan, Japanese, Korean, Dravidian and Indo-Pacific (e.g. Papua languages).
Austro-Asiatic	524	
Austronesian	2054	
Dravidian	542	
Indo-Pacific	29	
Japanese	2063	
Korean	4568	
Sino-Tibetan	13405 (34)	
Chinese	10304	
Tibeto-Burman	3067	
Thai	604	
Eskimo Aleut	631	This MP comprises all populations from Greenland, US, Canada and Russia, speaking one of the Eskimo-Aleut languages.
Eurasian	60966 (191)	This MP comprises all populations inhabiting the landmass between Atlantic, Indian and Pacific Ocean and speaking an Indo-European language.
Altaic	5226	
Caucasian	502	
European	48498 (5839)	On basis of R_{ST} analyses of the European populations sampled for the YHRD three main clusters emerged: the Western, Eastern and South-Eastern European MP.
Eastern European	12455	
South-Eastern European	5538	
Western European	24666	
Indian	2924	
Indo-Iranian	2803	
Uralic-Yukaghir	1013	
Native American	2151	This MP comprises all indigenous people who inhabit North, Central and South America since pre-Columbian times.
Admixed	15590	This MP comprises populations which members originate from two or more very distant source populations (e.g. the Mestizo populations of South America)

Table 2: Predefined Metapopulations

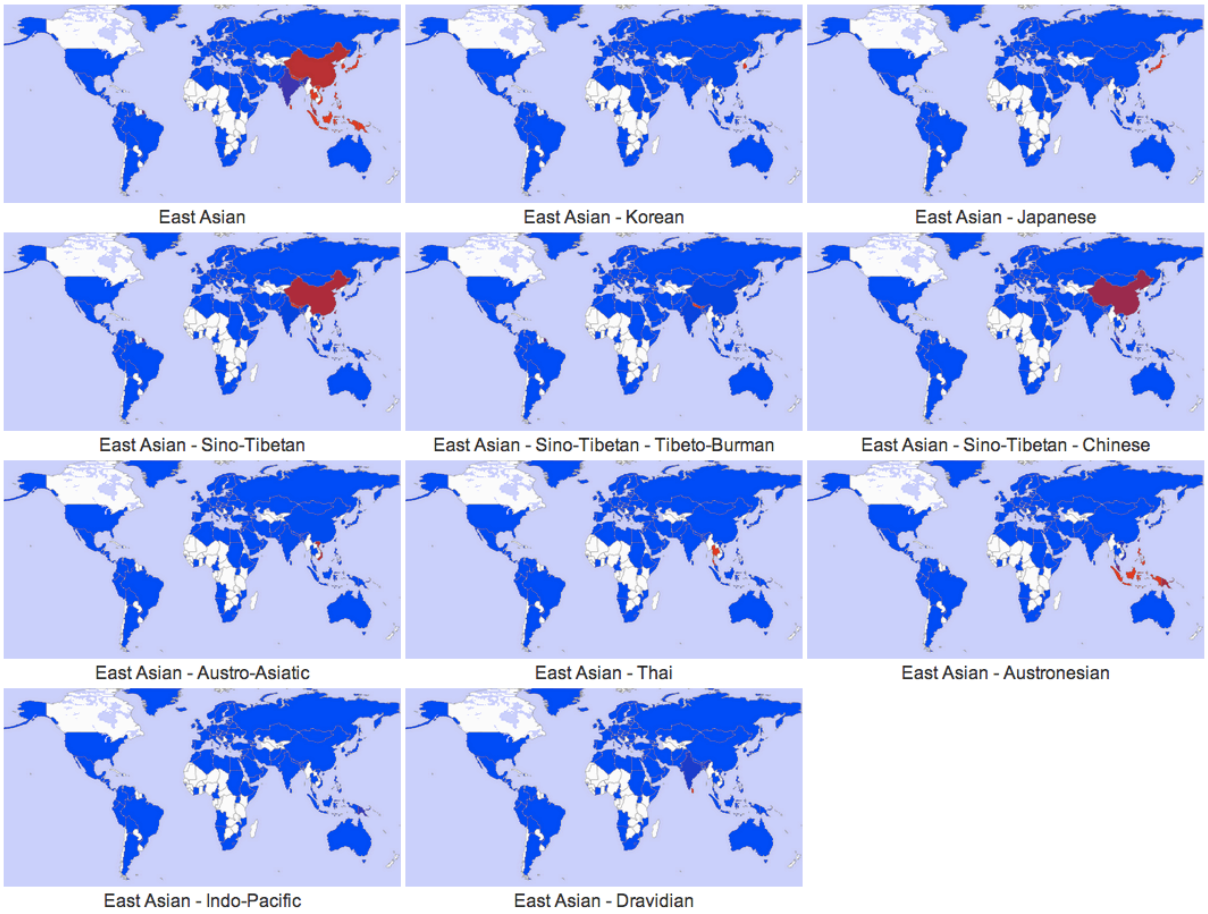


Figure 76: Distribution of East Asian Metapopulations

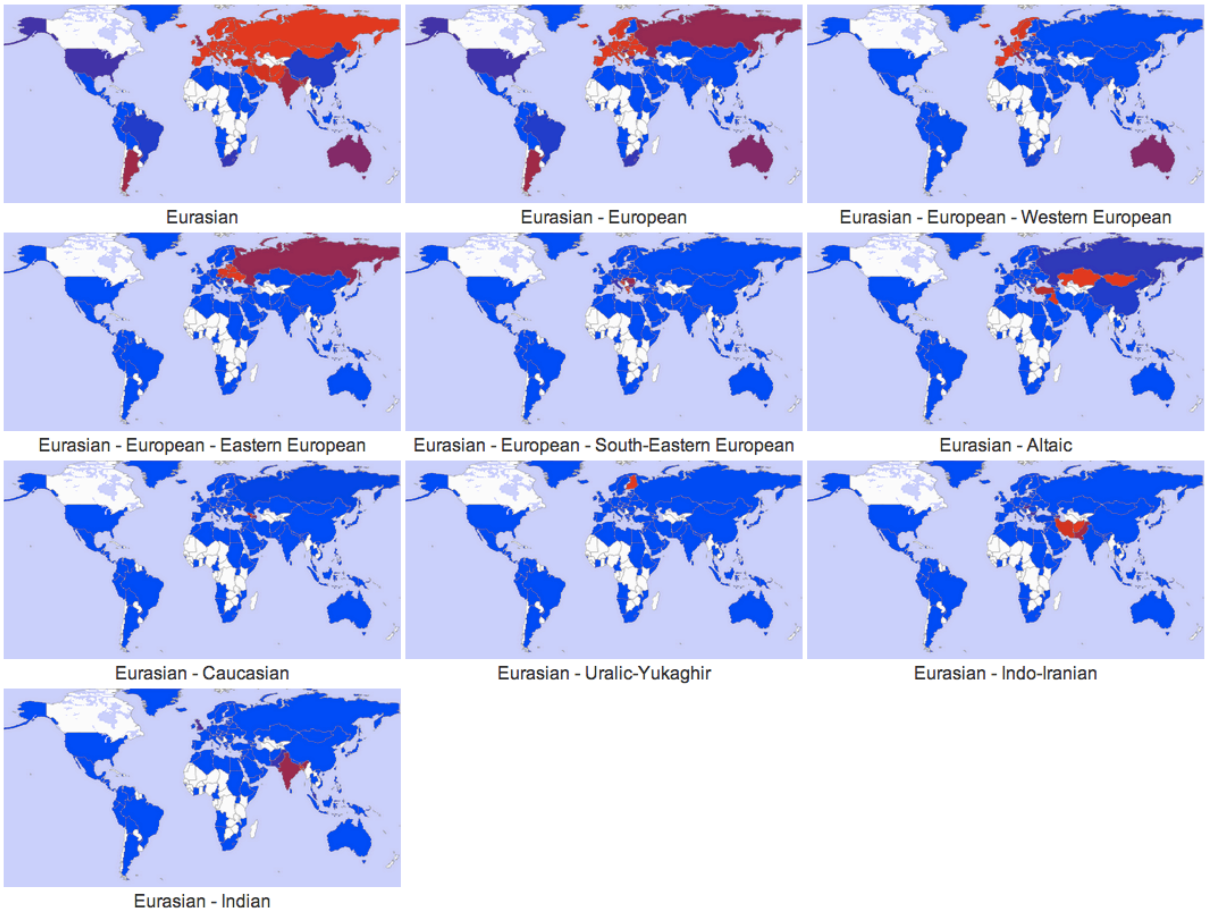


Figure 77: Distribution of Eurasian Metapopulations

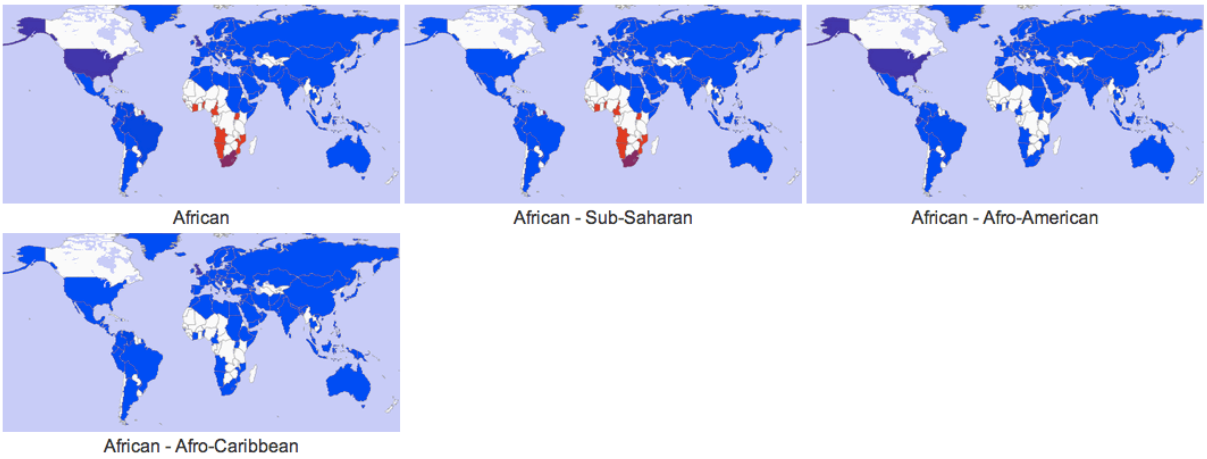


Figure 78: Distribution of African Metapopulations

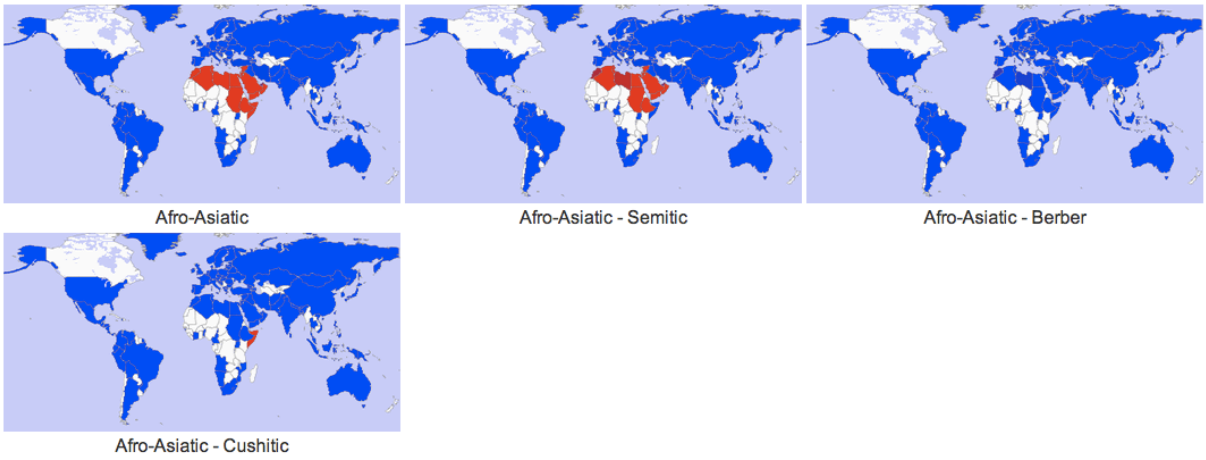


Figure 79: Distribution of Afro-Asiatic Metapopulations

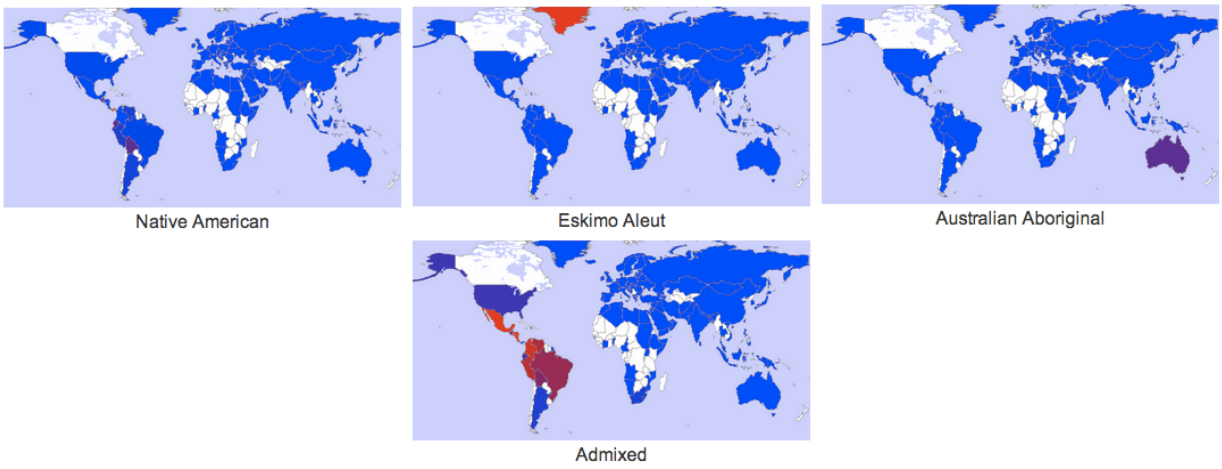


Figure 80: Distribution of Native-American, Eskimo-Aleut, Australian-Aboriginal and Admixed Metapopulations

4.1.4 Phylogenetic

The typing of Y chromosomes submitted to the YHRD is now continuously extended for binary Y-SNP polymorphisms. A grouping of Y-STR Haplotypes according to Haplogroups thus becomes feasible. The phylogeny of the Y chromosome defined by binary polymorphisms is well established and stable (Underhill et al. (2000), Hammer et al. (2001), Jobling and Tyler-Smith (2003) and Karafet et al. (2008)). All Y chromosomes sharing a mutation are related by descent, until a further mutation splits the branch. Haplotypes within a Haplogroup could be highly similar or even "identical by descent" (IBD). In thus, the Haplogroup could be used as a criterion to substructure the database according to the phylogenetic descent of samples. Even though the chronology of the SNP mutations is far less certain than the structure of the tree, many Haplogroups could be equated with events in human prehistory. The worldwide distribution of the patterns of the human Y-chromosome diversity has revealed clear geographically associated Haplogroups (Underhill et al. (2000)).

4.2 Haplogroups

A Haplogroup is defined by a specific Y-SNP mutation event. The YHRD favors a system of defining Y-chromosome haplogroups by letters A through to T, with further subdivisions using the Marker name.

4.3 Statistics

4.3.1 AMOVA (Analysis of Molecular Variance)

Derived from F-Statistics, AMOVA helps analyzing haploid data called phi-statistics. See Φ_{ST}, Φ_{CT} and Φ_{SC}. To assess the significance of Φ_{ST} values, permutational analysis of the null distribution drawn by randomly interchanging Haplotypes of populations is evaluated.
There are two different ways of dealing with a molecular distance of two Haplotypes: F_{ST}-based, where two Haplotypes are either equal ($F_{ST} = 0$) or have at least one inequality ($F_{ST} = 1$) and R_{ST}-based, where the sum of all squared differences is used (Excoffier et al. (1992) and Roewer et al. (1996)).

F_{ST} is the ratio of the average number of differences between pairs of chromosomes sampled within diploid individuals with the average number obtained when sampling chromosomes randomly from the population (excluding the grouping per individual).

$$F_{ST} = var(p)/P(q)$$

Φ_{ST} is describing the correlation between molecular diversity of random Haplotypes between two populations relative to random pairs of Haplotypes drawn from the whole species.

$$\Phi_{ST} = var(a) + var(b)/var(x)$$

Φ_{CT} is describing the correlation between molecular diversity of random Haplotypes within a group of populations relative to random pairs of Haplotypes drawn from the whole species.

$$\Phi_{CT} = var(a)/var(x)$$

Φ_{SC} is describing the correlation between molecular diversity of random Haplotypes within populations relative to random pairs of Haplotypes drawn from the region.

$$\Phi_{SC} = var(a)/var(a) + var(b)$$

4.3.2 F-Statistic

The values of F-statistics measure the correlation between genes drawn at different levels of a hierarchically subdivided population and allowing the characterization of the level of genetic distinctiveness of supposedly inbred or isolated populations and discrimination even between closely related populations, specifically the degree of (usually) a reduction in heterozygosity when compared to Hardy-Weinberg expectation.
F_{IT} is the inbreeding coefficient of an individual (I) relative to the total (T) population; F_{IS} is the inbreeding coefficient of an individual (I) relative to the sub-population (S), using the above for sub-populations and averaging them; and F_{ST} is the effect of sub-populations (S) compared to the total population (T). Those values are called *Fixation indices*.

4.3.3 MDS (Multidimensional scaling)

This analysis is used to assign pair wise similarities (or dissimilarities) to points in a N-dimensional space representing those similarities as distances between points. There are two major principles of MDS:

Metric MDS (M-MDS), a distance matrix D (similarities) into a set of coordinates such that the Euclidean distances derived from these coordinates fitting D as well as possible. The basic idea of M-MDS is to transform the distance matrix into a cross-product matrix and then to find its Eigen-decomposition which gives a principal component. This requires linearity assumption to be met.

Non-Metric MDS (N-MDS) on the other hand, uses the rank of a distance matrix delta (dissimilarities) to iteratively assigning locations to monotonic parts of delta. In every iteration, the configuration of assigned locations evaluated with respect to a stress criterion (how well does the configuration approximate the original input dissimilarities).

4.3.4 PCA (Principal component analysis)

PCA is useful to reduce the dimension of the input data, which may be highly correlated (redundant). This technique has three effects: it orthogonalizes the components of the input vectors (so that they are uncorrelated with each other), it orders the resulting orthogonal components (principal components) so that those with the largest variation come first, and it eliminates those components that contribute the least to the variation in the data set. In a result there are two principal components: The first is a single axis in space. When you project each observation on that axis, the resulting values form a new variable. And the variance of this variable is the maximum among all possible choices of the first axis. The second is another axis in space, perpendicular to the first. Projecting the observations on this axis generates another new variable. The variance of this variable is the maximum among all possible choices of this second axis.

5 References

References

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data., *Genetics* **131**(2): 479–491.

Gusmão, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., Morling, N., Prinz, M., Roewer, L., Tyler-Smith, C., Schneider, P. M. and International Society of Forensic Genetics (2006). DNA commission of the international society of forensic genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis., *Int J Legal Med* **120**(4): 191–200.

Hammer, M. F., Karafet, T. M., Redd, A. J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H. and Zegura, S. L. (2001). Hierarchical patterns of global human y-chromosome diversity., *Mol Biol Evol* **18**(7): 1189–1203.

Hanski, I. and Gilpin, M. (1997). *Metapopulation Biology: Ecology, Genetics, and Evolution.*, Academic Press, San Diego.

Jobling, M. A. and Tyler-Smith, C. (2003). The human y chromosome: an evolutionary marker comes of age., *Nat Rev Genet* **4**(8): 598–612.

Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L. and Hammer, M. F. (2008). New binary polymorphisms reshape and increase resolution of the human y chromosomal haplogroup tree., *Genome Res* **18**(5): 830–838.

Roewer, L. (2009). Y chromosome STR typing in crime casework., *Forensic Sci Med Pathol* **5**(2): 77–84.

Roewer, L., Croucher, P. J. P., Willuweit, S., Lu, T. T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M. A., Tyler-Smith, C. and Krawczak, M. (2005). Signature of recent historical events in the european y-chromosomal STR haplotype distribution., *Hum Genet* **116**(4): 279–291.

Roewer, L., Kayser, M., de Knijff, P., Anslinger, K., Betz, A., Caglià, A., Corach, D., Füredi, S., Henke, L., Hidding, M., Kärgel, H. J., Lessig, R., Nagy, M., Pascali, V. L., Parson, W., Rolf, B., Schmitt, C., Szibor, R., Teifel-Greding, J. and Krawczak, M. (2000). A new method for the evaluation of matches in non-recombining genomes: application to y-chromosomal short tandem repeat (STR) haplotypes in european males., *Forensic Sci Int* **114**(1): 31–43.

Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., Bakker, E., Krawczak, M. and de Knijff, P. (1996). Analysis of molecular variance (AMOVA) of y-chromosome-specific microsatellites in two closely related human populations., *Hum Mol Genet* **5**(7): 1029–1033.

Underhill, P. A., Shen, P., Lin, A. A., Jin, L., Passarino, G., Yang, W. H., Kauffman, E., Bonné-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J. R., Mehdi, S. Q., Seielstad, M. T., Wells, R. S., Piazza, A., Davis, R. W., Feldman, M. W., Cavalli-Sforza, L. L. and Oefner, P. J. (2000). Y chromosome sequence variation and the history of human populations., *Nat Genet* **26**(3): 358–361.

Willuweit, S., Roewer, L. and The International Forensic Y Chromosome User Group (2007). Y chromosome haplotype reference database (YHRD): update., *Forensic Sci Int Genet* **1**(2): 83–87.

Wolf, A., Caliebe, A., Junge, O. and Krawczak, M. (2005). Forensic interpretation of y-chromosomal DNA mixtures., *Forensic Sci Int* **152**(2-3): 209–213.