

dbGaP's New Ancestry Composition Visualization tool and GRAF Software

- Closed captioning: www.captionedtext.com & enter 3639837
- The **recording of this webinar** will be on our YouTube channel in the Webinars playlist in a few days: [youtube>/user/NCBINLM/playlists](https://www.youtube.com/user/NCBINLM/playlists)
- Type in the **Questions Pod** to ask questions when you think of them. *Don't wait until the end.*
- Answers to all questions will be available after the webinar with a link on our Webinars page: [ncbi>/home/coursesandwebinars.shtml](http://ncbi/home/coursesandwebinars.shtml)
- Slides, Q&A: <https://go.usa.gov/xQEwa>

<ncbi> = www.ncbi.nlm.nih.gov
<youtube> = www.youtube.com



U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI Webinars



dbGaP's New Ancestry Composition Visualization Tool and GRAF Software

Wayne Matten – presenter
matten@ncbi.nlm.nih.gov

Yumi (Jimmy) Jin - developer
jinyu@ncbi.nlm.nih.gov



U.S. National Library of Medicine
National Center for Biotechnology Information

GRAF – Genetic Relationship and Fingerprinting

- Introduce GRAF
 - Why was it created?
 - How does it work?
- Demonstrate how GRAF can help you
 - Identify duplicates and closely related subjects
 - Determine subject ancestries

Why GRAF was Created

- Originally a curation tool for dbGaP
 - detect duplicates and close relatives *across studies*
 - check the subject-sample mapping (SSM) and pedigree files against submitted dbGaP genotypes
- Expanded to determine subject ancestries

Computed Ancestry Links

dbGaP Advanced Search ?

Save Query

Type Keyword or Phrase

Show All Filters

Study Disease/Focus (482)

Study Design (15)

Study Molecular Data Type (37)

Study Markerset (159)

NIH Institute (22)

Study Consent (457)

Study Type (67)

Study Subject Count

Studies (1092) Variables (301212) Phenotype Datasets (6596) Documents (7789)

Molecular Datasets (953) Analyses (4540)

1 [NEI Age-Related Eye Disease Study \(AREDS\)](#)

Accession	phs000001.v3.p1
Study Disease/Focus	Cataract
Study Design	Case-Control
Study Markerset	Illumina100K, Affymetrix100K, HumanOmni2.5-4v1_D
Study Molecular Data Type	Not Provided
Study Content	16 phenotype datasets, 35 variables, 48 documents, 3 analyses, 3 molecular datasets, 4757 subjects, 6962 samples, 1 sub-studies
Ancestry (computed)	population graph European (575), African American (11), Hispanic1 (2), South Asian (2), Other (3)
NIH Institute	NEI
Study Consent	EDO — Eye disease research only , GRU — General research purposes
Release Date	2012-04-05
Embargo Release Date	2016-02-11

The Age-Related Eye Disease Study (AREDS) was initially designed as a long-term multi-center, prospective study of the clinical course of age-related macular degeneration (AMD) and age-related cataract. In addition ... to collecting natural history data, AREDS included a clinical trial of high-dose vitamin and mineral supplements for AMD and a clinical trial of

2 [Framingham Cohort](#)

Accession	phs000007.v29.p10
Study Disease/Focus	Cardiovascular Diseases
Study Design	Prospective Longitudinal Cohort
Study Markerset	HuEx-1_0-st, custom_probe_set, Affymetrix_50K, Affymetrix_100K, Genome-Wide_Human_SNP_Array_5.0, LegacySet, HapMap_phaseII, CVDSNP55v1_A, exome_vcf_grc37, HumanOmni5-4v1_B, targeted_region_sequencing, SNP-PCR, HumanMethylation450, WES, methylation_qc38



Computed Ancestry Links



National Eye Institute (NEI) Age-Related Eye Disease Study (AREDS)

dbGaP Study Accession: phs000001.v3.p1

Request Access

-
- Study Weblinks: [AREDS, The National Eye Institute](#); [AREDS, The EMMES Corporation](#)
 - Study Type: Case-Control
 - dbGaP estimated [ancestry components](#) using [GRAF-pop](#)
 - Number of study subjects that have individual level data available through Authorized Access: 4757
 - 4757 phenotyped subjects

Ancestry Component Using GRAF-pop

Click the checkboxes to select/unselect study-reported populations:

Study-Reported Population	
WHITE - NOT OF HISPANIC ORIGIN	<input checked="" type="checkbox"/>
UNSPECIFIED	<input checked="" type="checkbox"/>
BLACK, NOT OF HISPANIC ORIGIN	<input checked="" type="checkbox"/>
OTHER	<input checked="" type="checkbox"/>

Set graph parameters:

width dot size min #SNPs rotation

min max hide unselected plot

GD1 GD1 GD4

min max min GD4 max

GD2 GD2 GD4

Color/Group subjects by: population values

Select study-reported populations to show the areas that include 95% dbGaP subjects using color :

- European/White/Caucasian
- East Asian
- Puerto Rican/Dominican
- Asian/Pacific Islander
- African (Ghana/Yoruba)
- African American/Black
- Mexican/Latino
- Asian
- Indian/Pakistani

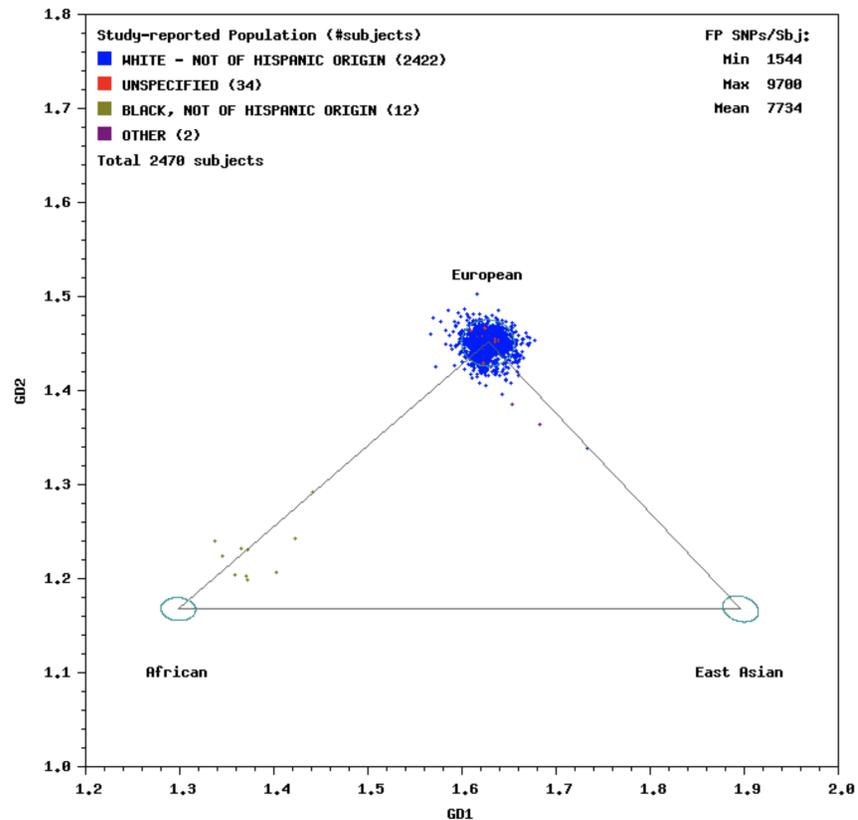


U

N:

- (Chinese/Japanese)
- Asian/Pacific Islander

phs000001.v3 Populations Determined Using Fingerprint SNPs



How does GRAF Work?

Identify duplicate and closely related subjects

1. Genotypes of 10,000 *preselected* fingerprint SNPs (FP SNPs)
2. Two statistical metrics
 - All genotype mismatch rate (AGMR)
 - all FP SNPs with genotypes are counted
 - Homozygous genotype mismatch rate (HGMR)
 - only FP SNPs where both samples are homozygous are counted

How does GRAF Work?

Infer subject ancestry, GRAF-pop

1. Cluster subjects using genetic distances (GD) to several reference populations
2. Project clusters onto a fixed plane and estimate ancestral proportions for each subject from barycentric coordinates
3. Distinguish subjects with four statistical metrics, GD1, GD2, GD3, and GD4
4. View all subjects on the same scatter plot

How Can GRAF Help You?

1. Download and run GRAF

- your own datasets
 - analysis of a new subjects – did they participate in an existing study?
- dbGaP or other datasets

2. dbGaP submitter

- check for potential errors
- curators will send you GRAF results based on your datasets

Where to Get GRAF



Software

- **GRAF:** GRAF (**Genetic Relationship and Fingerprinting**) is a C++ program that quickly finds the closely related subjects using SNP genotype data.
 - The program can be used to validate the relationships of subjects and samples reported in the pedigree file and the subject-sample mapping file (SSM) against the genotype data.
 - The program can also be used to determine the ancestry of the subjectsFor a more detailed description, see GRAF_ReadMe.docx in the package.

Click the following link to [download GRAF](#). (Latest version: 2.3. Last updated: 04/17/2018)

REFERENCE: Jin Y, Schäffer AA, Sherry ST, and Feolo M (2017). Quickly identifying identical and closely related subjects in large databases using genotype data. *PLoS One*. 12(6):e0179106. [\[Abstract\]](#) [\[PDF\]](#)

- **TransEAV:** dbGaP requires that the submitted phenotypic datasets be rectangular tables with each row representing one subject or sample, and each column representing a phenotypic trait or attribute (called variable in dbGaP), and each cell storing one attribute value. However, sometimes datasets are collected and recorded using Entity-Attribute-Value model (EAV) model. In EAV model, one dataset table usually has three columns: subject (or sample), attribute, and value, and each row stores only one attribute value for one subject or sample. This script converts a dataset in EAV model to a rectangular table that can be submitted to dbGaP. For a more detailed description, please see README.txt in the package.

GRAF Package Contents

GRAF_ReadMe_20180417.docx
GRAF-popDocumentation_20180417.docx
PlotGraf.pl
PlotPopulations.pl
graf
graf_dups
FP_SNPs.txt
SsToRs.txt

programs

— Fingerprinting SNPs

perlegen_hapmap.bed
perlegen_hapmap.bim
perlegen_hapmap.fam

G1000FpGeno.bed
G1000FpGeno.bim
G1000SbjPop.txt
G1000SbjSuperPop.txt
G1000_sbj_list.txt

affy_hapmap.bed
affy_hapmap.bim
affy_hapmap.fam
affy_hapmap_fake_pedigree.txt
affy_hapmap_fake_ssm.txt

Example files

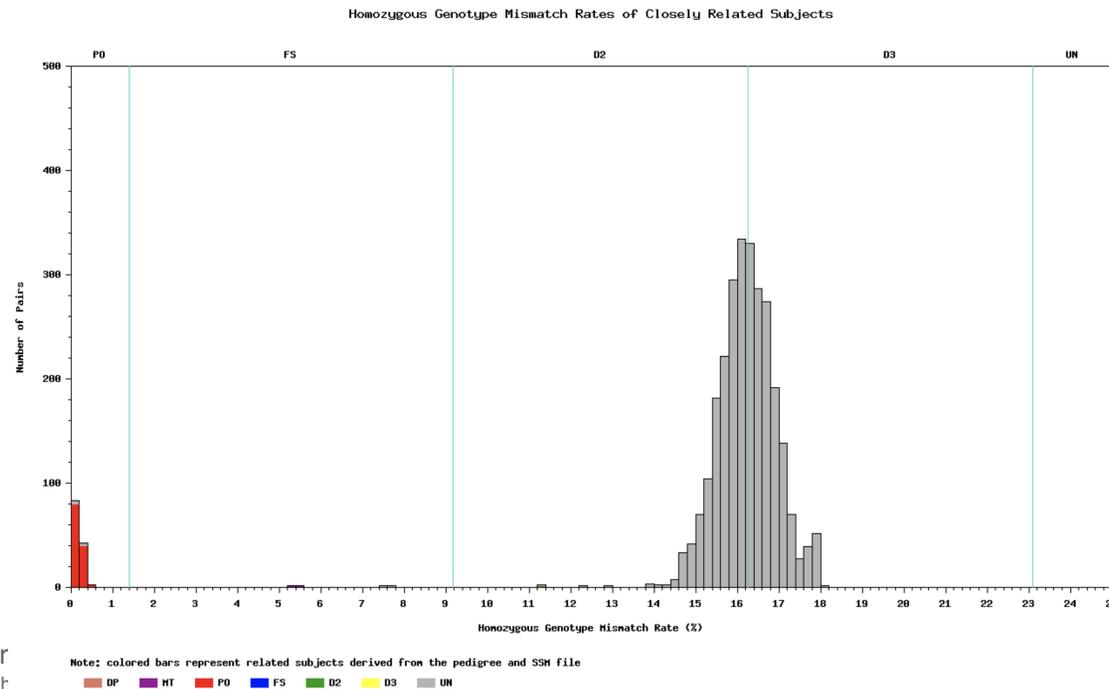
GRAF Command line

```
$ graf -plink affy_hapmap -out aff_hapmap_rels.txt
```

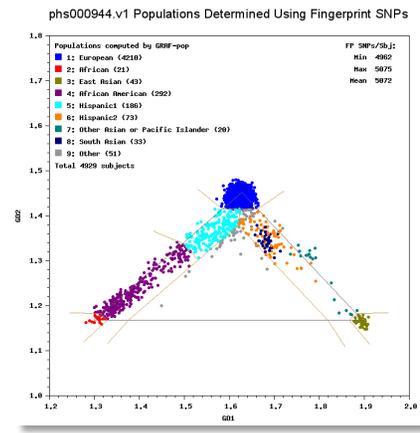
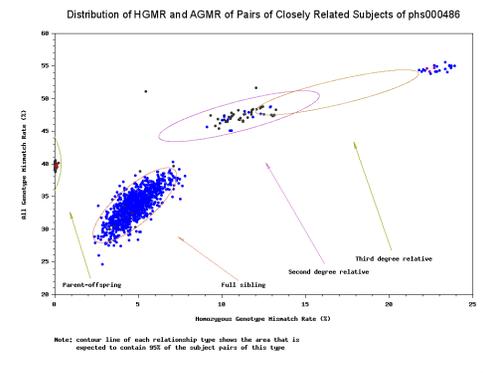
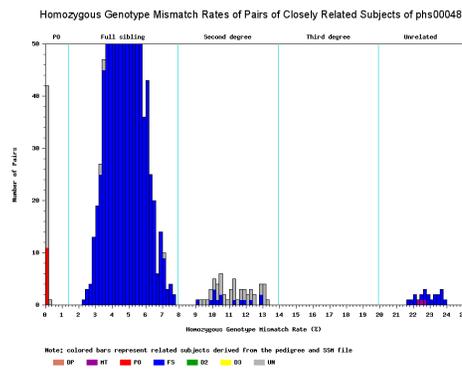
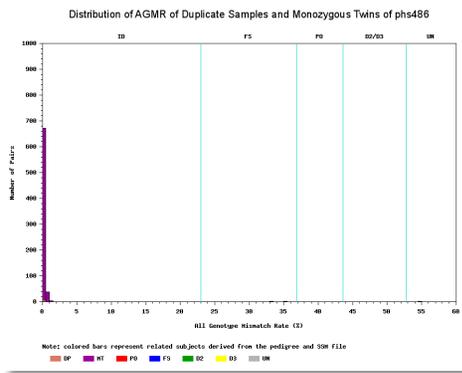
```
affy_hapmap.bed }  
affy_hapmap.bim } PLINK set  
affy_hapmap.fam }  
  
affy_hapmap_fake_pedigree.txt  
affy_hapmap_fake_ssm.txt  
affy_hapmap_ssm.txt  
comb_hapmap_ssm.txt
```

GRAF Command line

```
$ PlotGraf.pl aff_short_rels.txt aff_short_rels_scatter.png
```

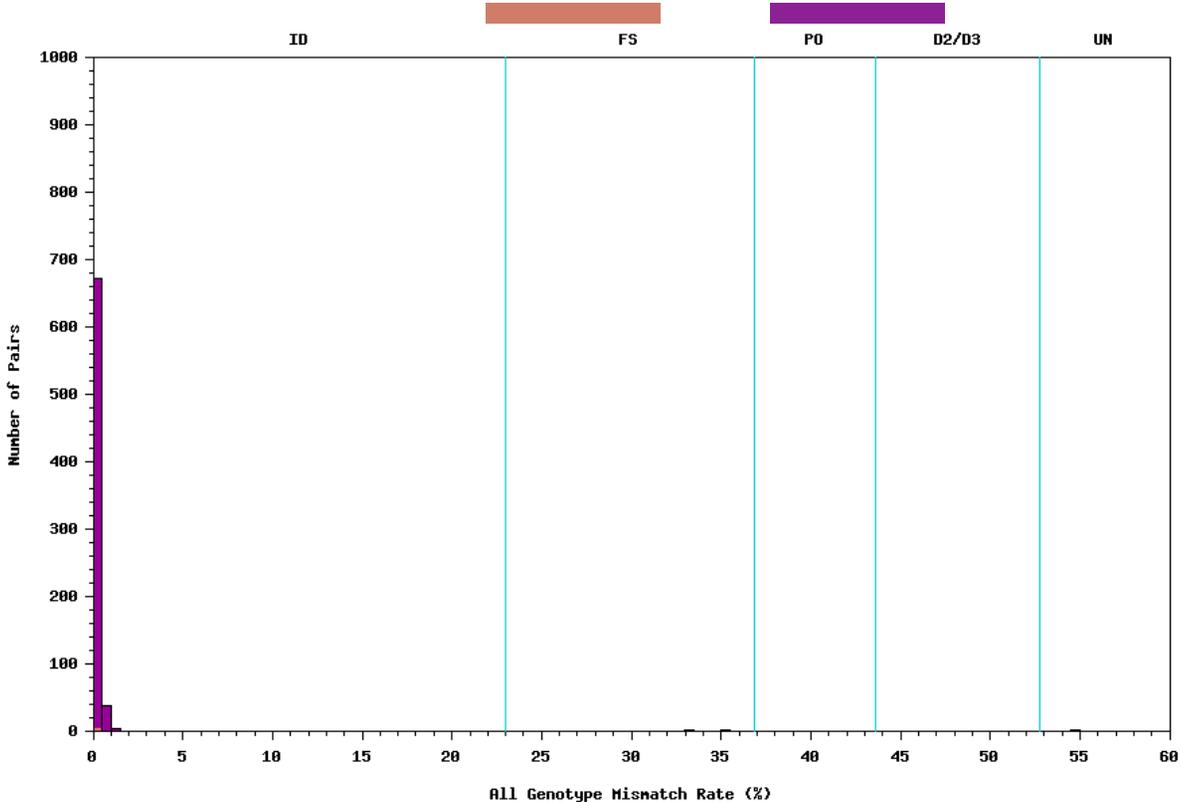


GRAF Results



GRAF Results - AMGR

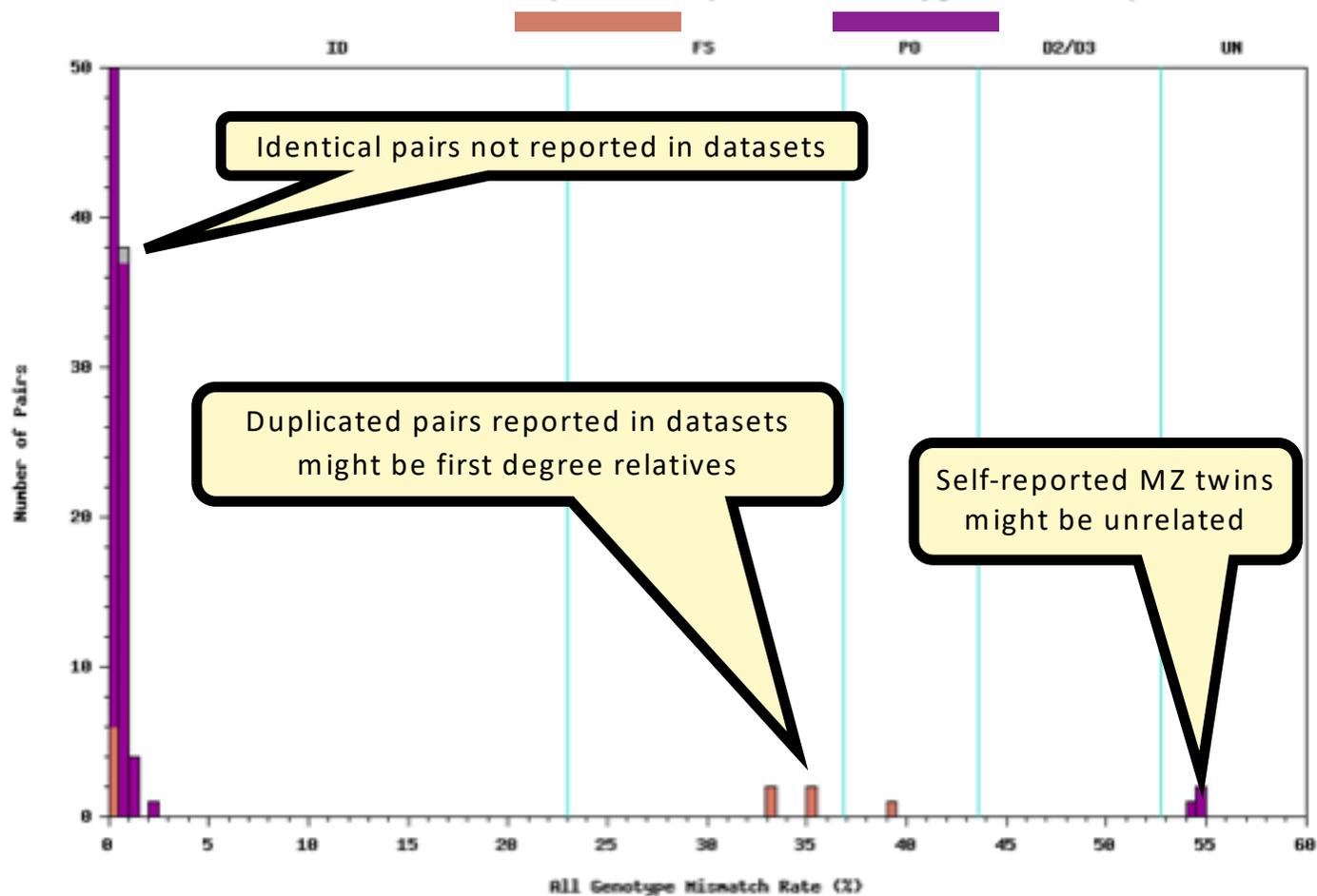
Distribution of AGMR of Duplicate Samples and Monozygous Twins of phs486



Y-axis: Number of Pairs

X-axis: AMGR

Distribution of AGMR of Duplicate Samples and Monozygous Twins of phs486

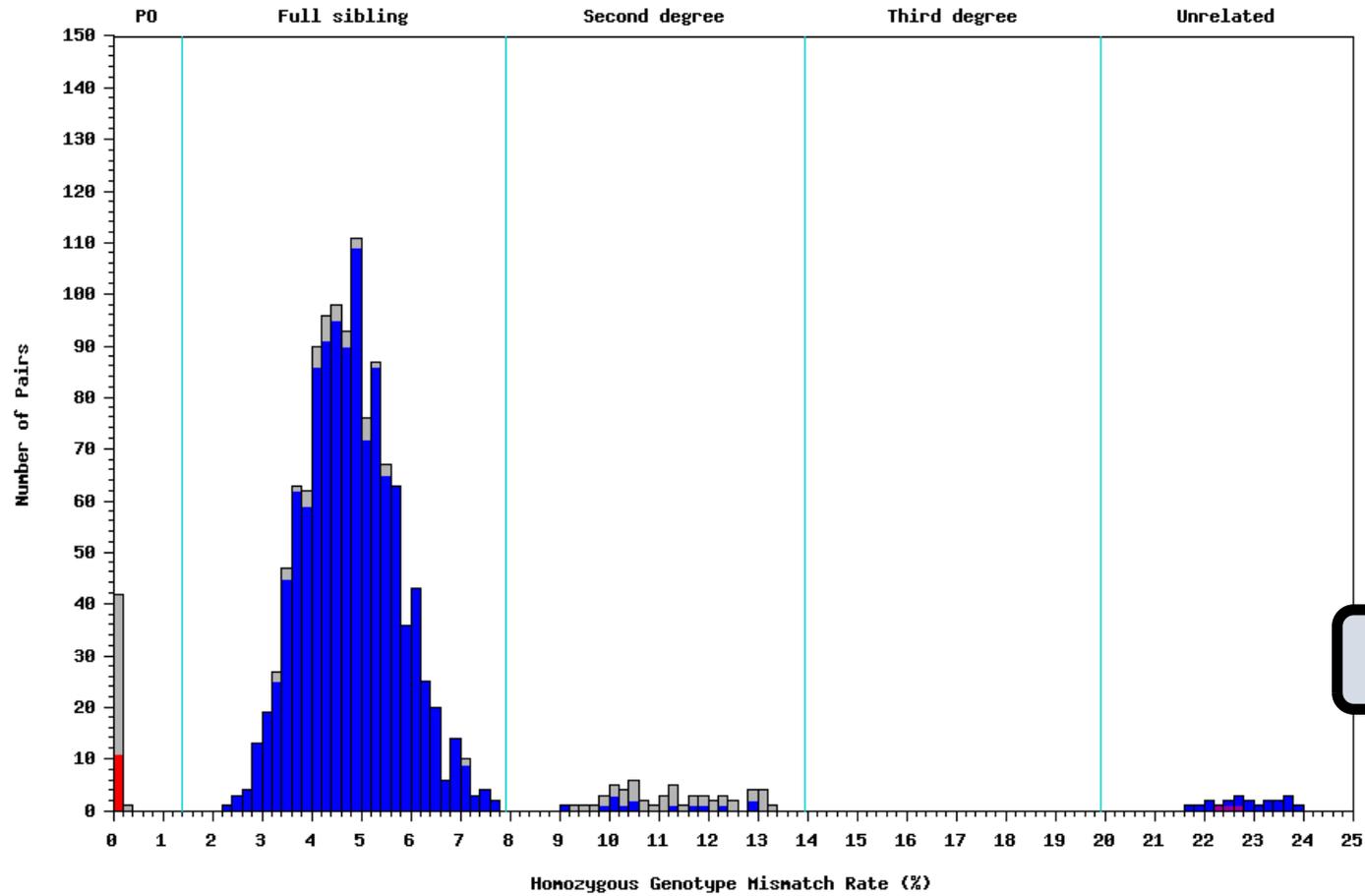


Identical pairs not reported in datasets

Duplicated pairs reported in datasets might be first degree relatives

Self-reported MZ twins might be unrelated

Homozygous Genotype Mismatch Rates of Pairs of Closely Related Subjects of phs000486

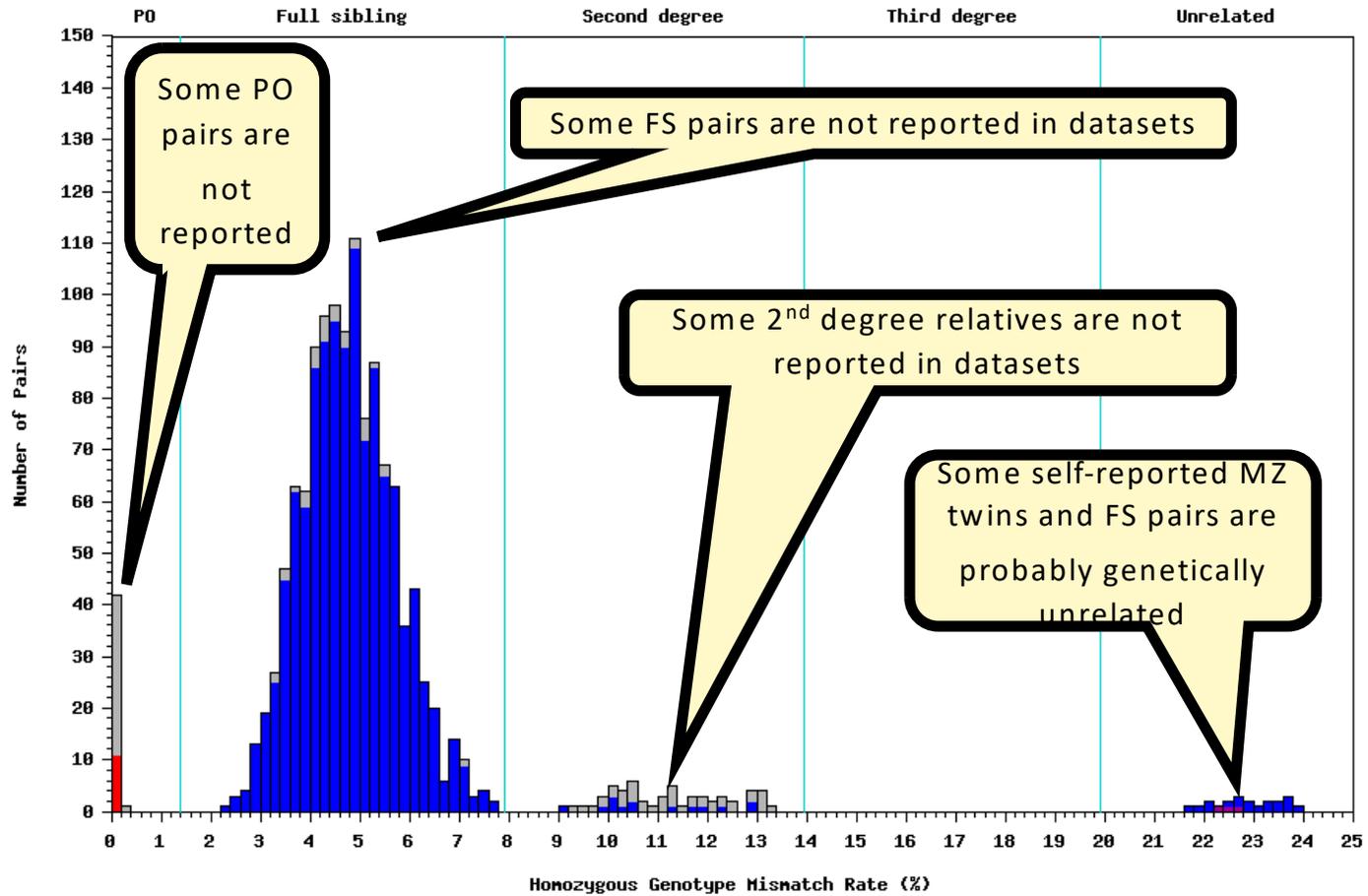


X-axis: HMGR

Note: colored bars represent related subjects derived from the pedigree and SSM file

- DP
- HT
- P0
- FS
- D2
- D3
- UN

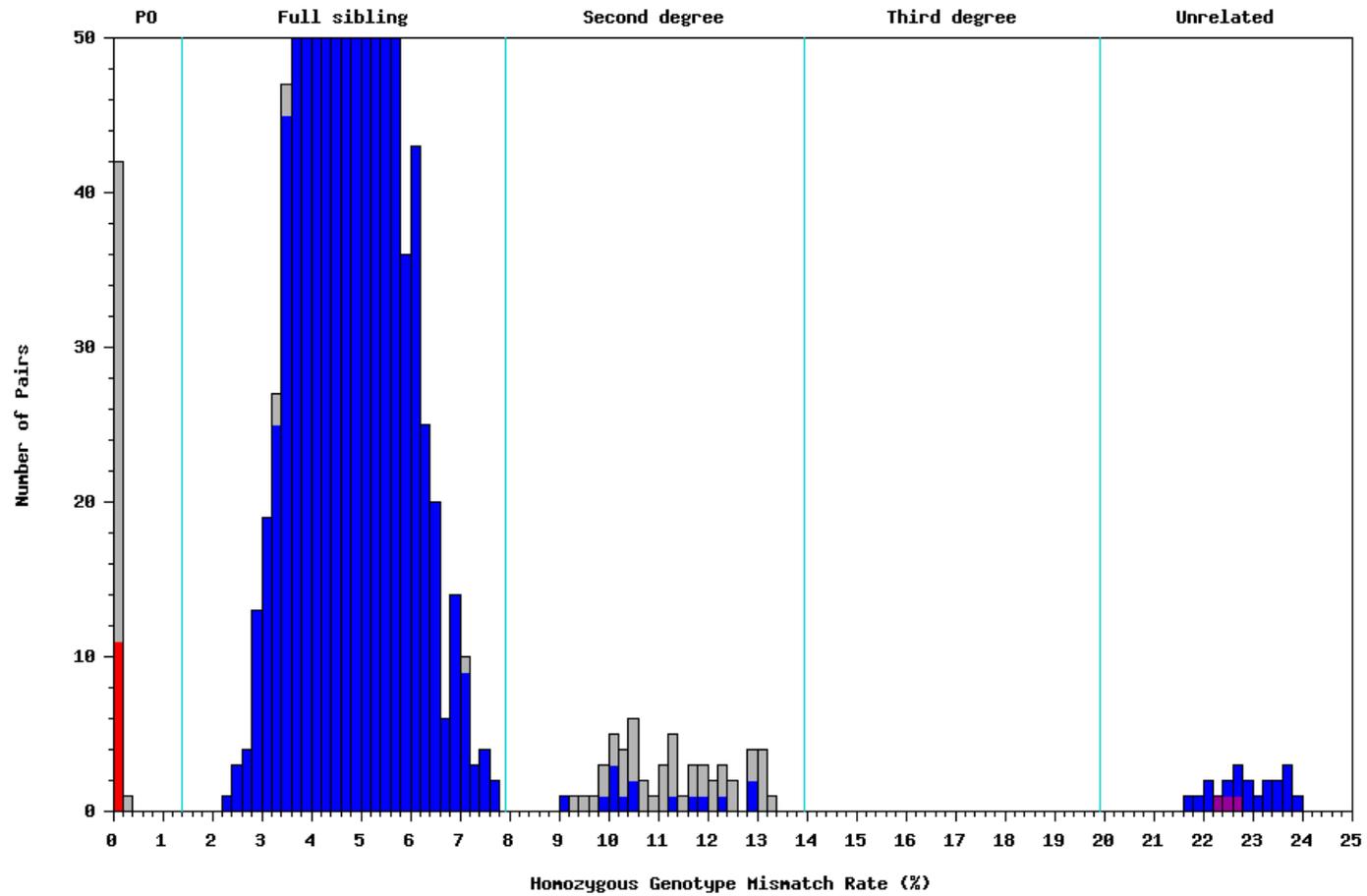
Homozygous Genotype Mismatch Rates of Pairs of Closely Related Subjects of phs000486



Note: colored bars represent related subjects derived from the pedigree and SSM file

DP HT PO FS D2 D3 UN

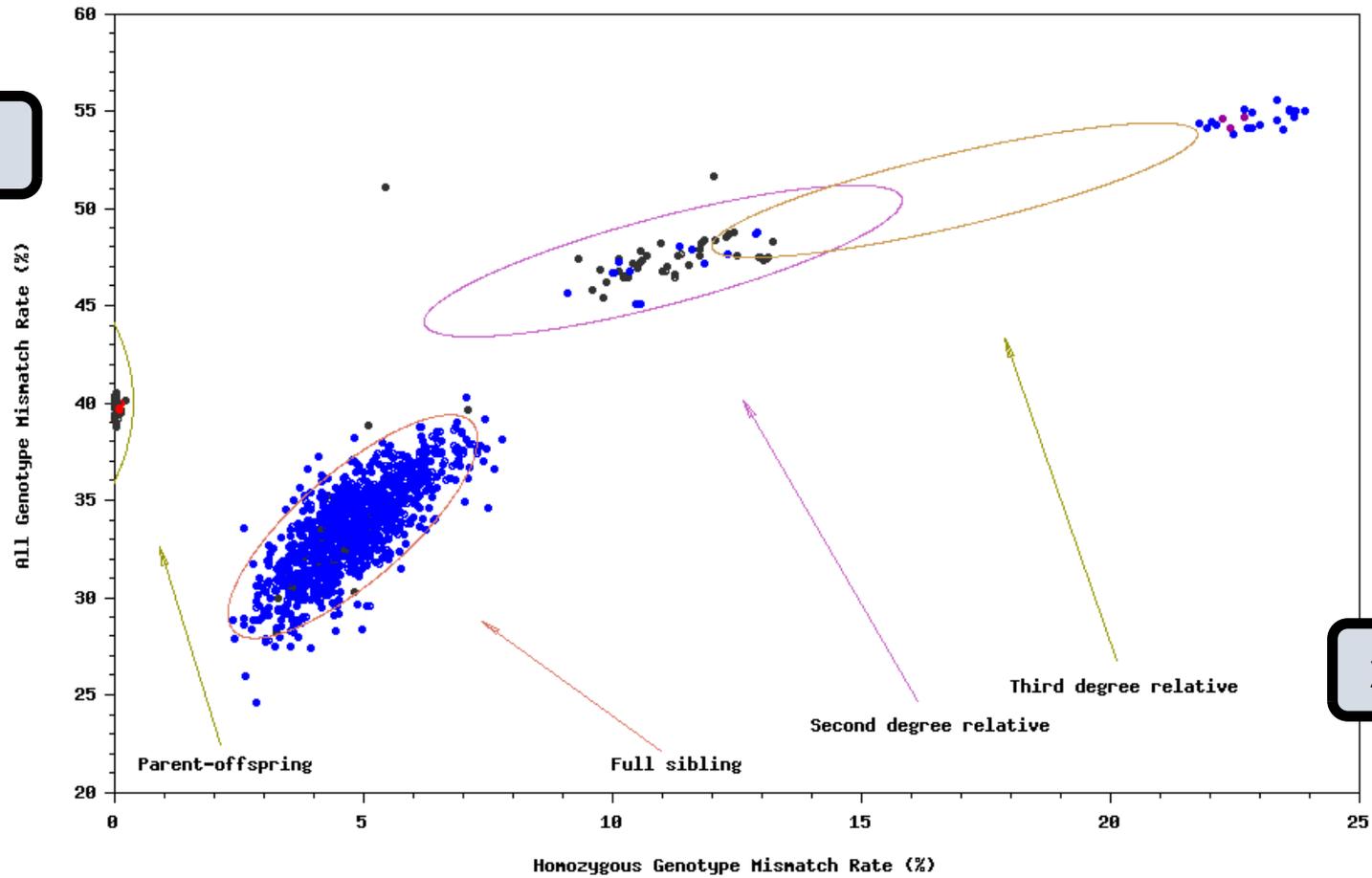
Homozygous Genotype Mismatch Rates of Pairs of Closely Related Subjects of phs000486



Note: colored bars represent related subjects derived from the pedigree and SSM file

DP HT P0 FS D2 D3 UN

Distribution of HGMR and AGMR of Pairs of Closely Related Subjects of phs000486

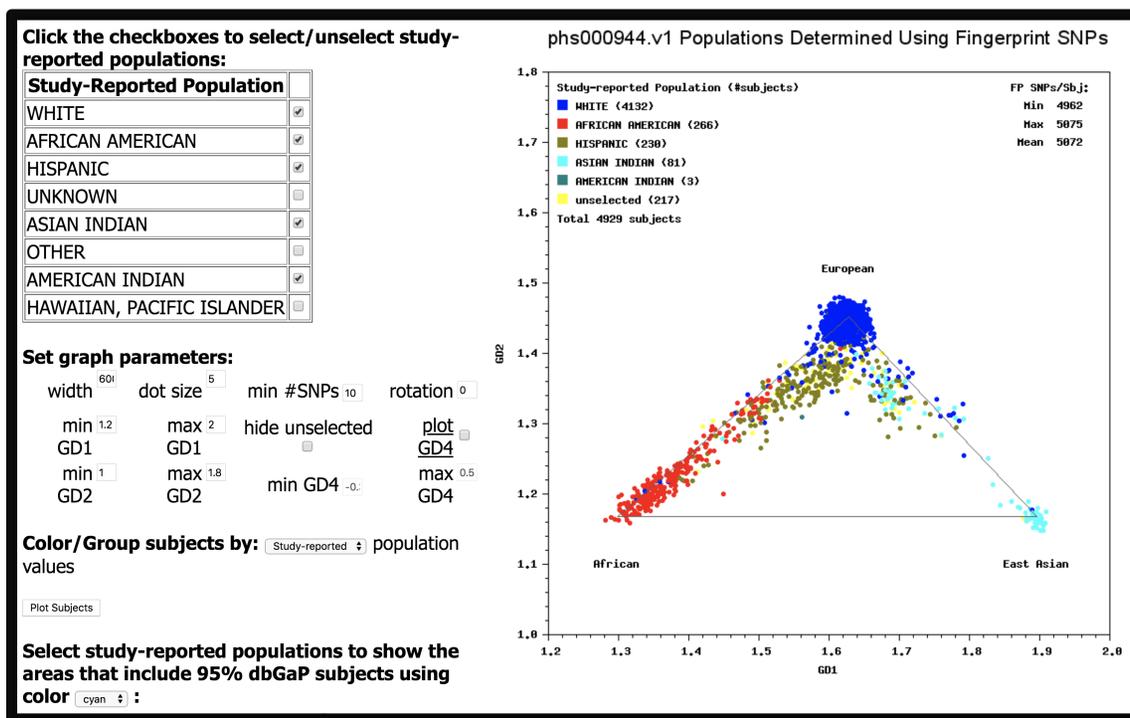


Y-axis: AMGR

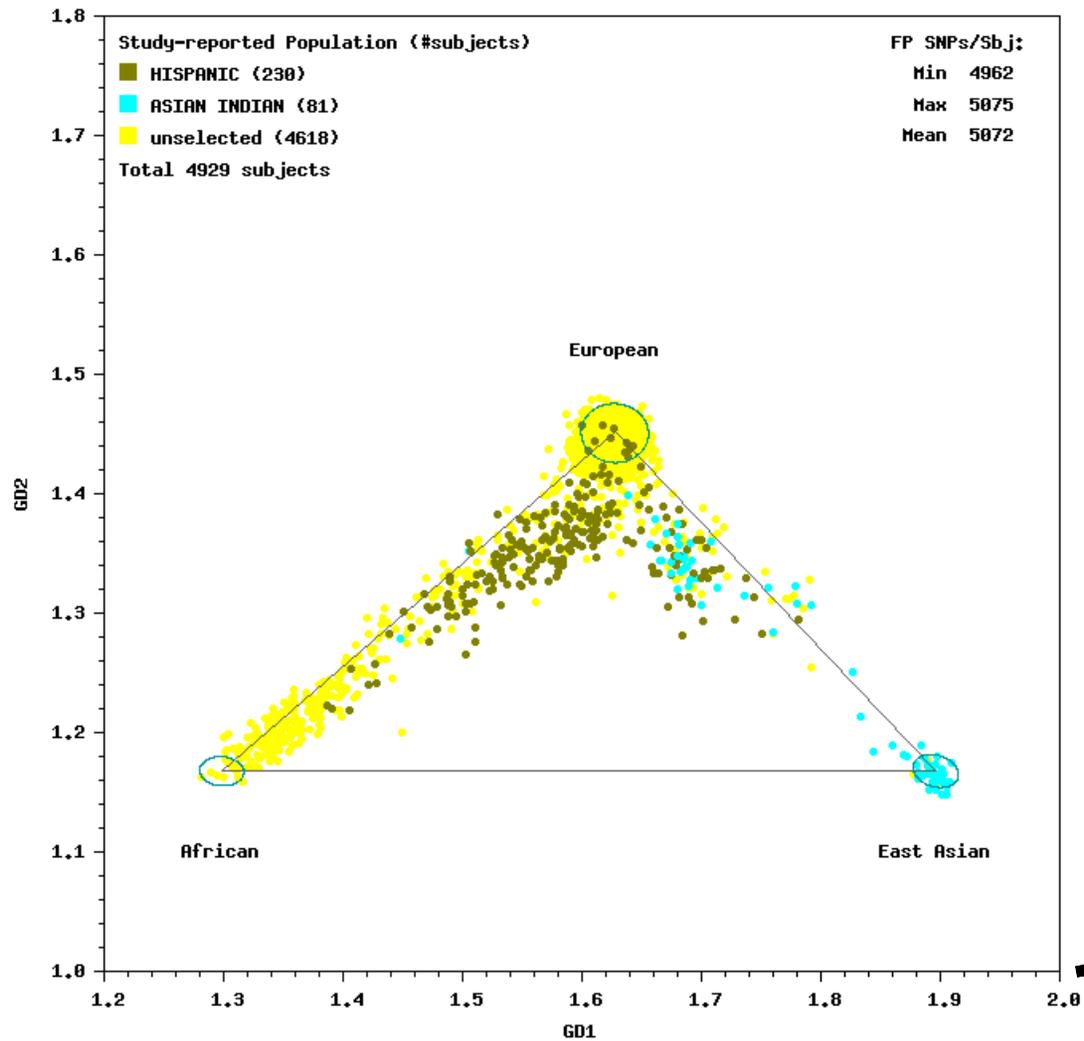
X-axis: HMGR

Note: contour line of each relationship type shows the area that is expected to contain 95% of the subject pairs of this type

GRAF-pop, Infer Subject Ancestry



phs000944.v1 Populations Determined Using Fingerprint SNPs



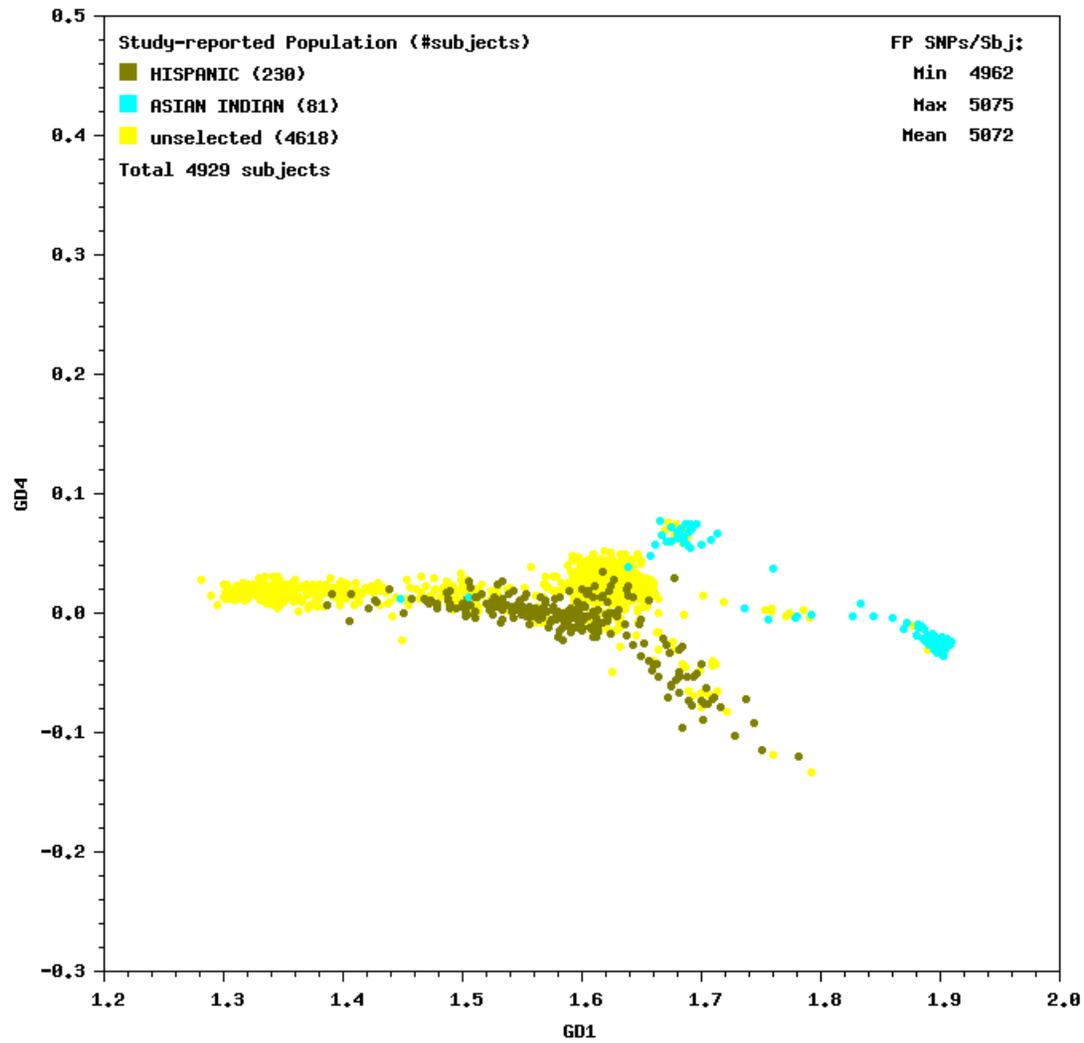
Y-axis: GD2

X-axis: GD1

phs000944.v1 Populations Determined Using Fingerprint SNPs

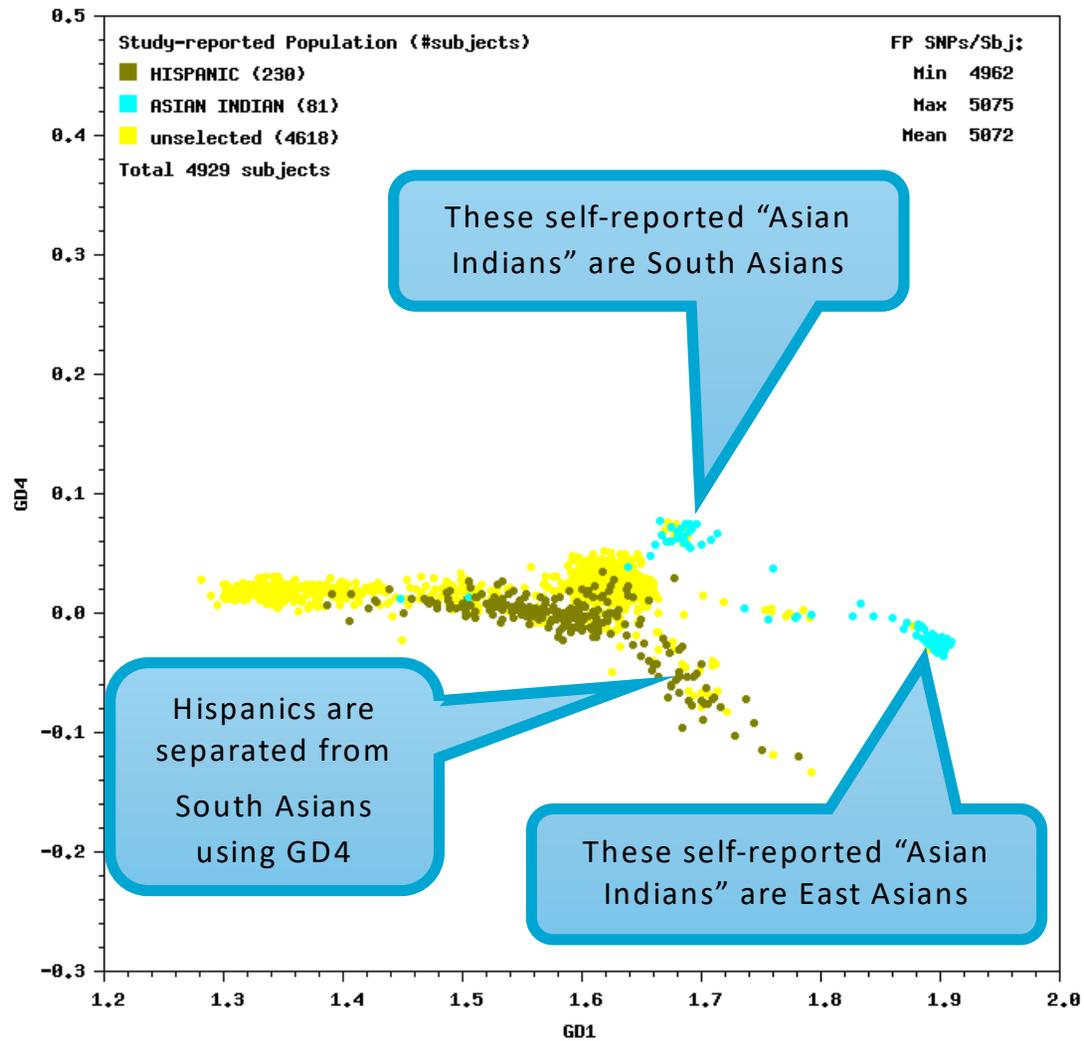


phs000944.v1 Populations Determined Using Fingerprint SNPs

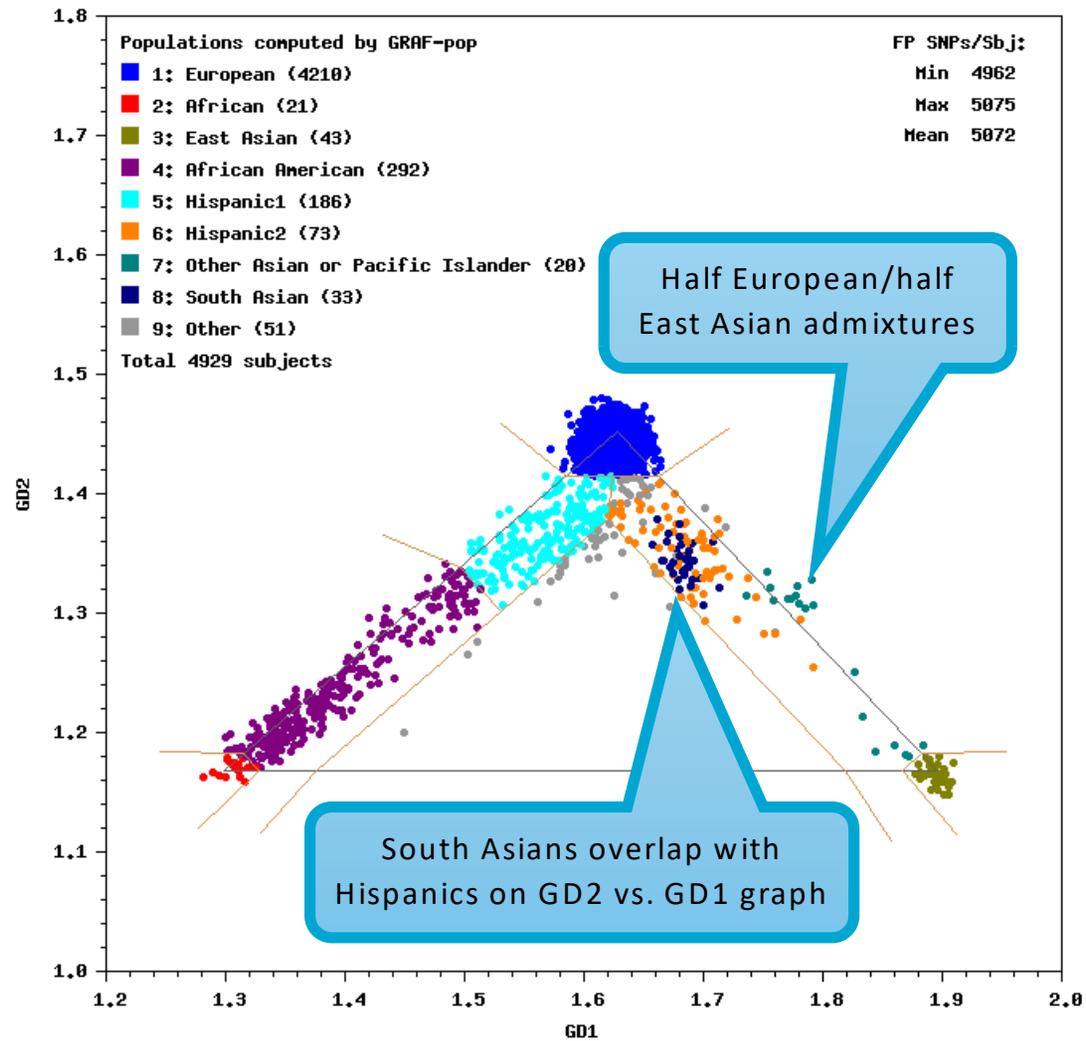


Y-axis: GD4

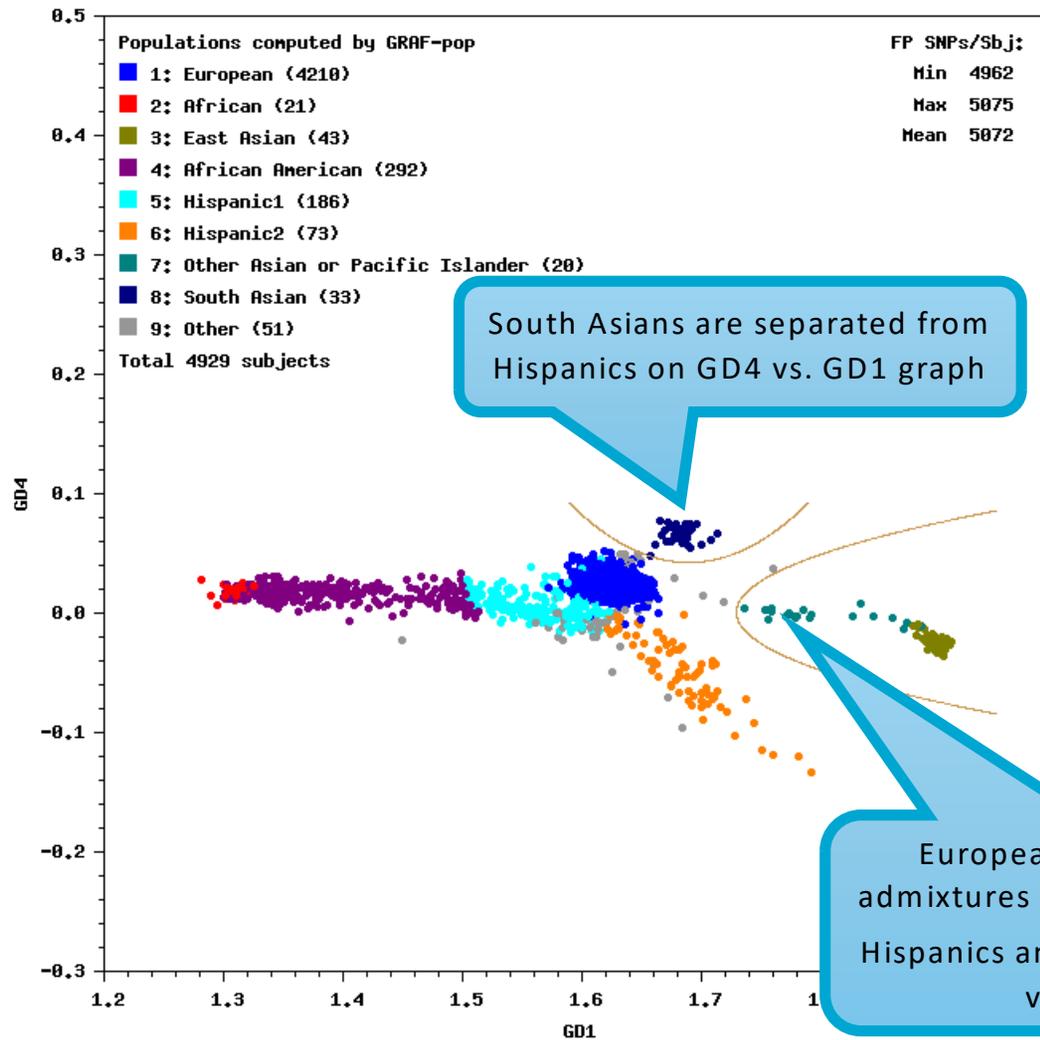
phs000944.v1 Populations Determined Using Fingerprint SNPs



phs000944.v1 Populations Determined Using Fingerprint SNPs



phs000944.v1 Populations Determined Using Fingerprint SNPs



View on Web Page

- Relatedness scatter plot (HGMR+AGMR)
- GRAF-pop, GD1 vs GD4
- GRAF-pop, select a region

Questions?

Places to Learn More

- Download GRAF software: www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/Software.cgi
- Materials: www.ncbi.nlm.nih.gov/home/coursesandwebinars/
- NCBI Announcements: <https://ncbiinsights.ncbi.nlm.nih.gov/ncbi-rss-feeds-listservs/>
- Factsheets: <ftp://ncbi.nih.gov/pub/factsheets/>
- NCBI YouTube Channel: www.youtube.com/ncbinlm

For help with NCBI resources:

info@ncbi.nlm.nih.gov

For questions about webinars:

webinars@ncbi.nlm.nih.gov



U.S. National Library of Medicine
National Center for Biotechnology Information

NCBI Webinars

