ACMG-ClinVar Penetrance RMarkdown

James Diao, under the supervision of Arjun Manrai

April 12, 2017

Contents

Working Directory: /Users/jamesdiao/Documents/Kohane_Lab/2017-ACMG-penetrance/ACMG_Penetrance

1 Download, Transform, and Load Data

1.1 Collect ACMG Gene Panel

http://www.ncbi.nlm.nih.gov/clinvar/docs/acmg/

```
## Table from ACMG SF v2.0 Paper 60 x 8 (selected rows):
```

	Phenotype	MIM_disorder	PMID_Gene_Reviews_entry
N1	Hereditary breast and ovarian cancer	604370|612555	20301425
N2	Hereditary breast and ovarian cancer	604370|612555	20301425
N3	Li-Fraumeni syndrome	151623	20301488
N4	Peutz-Jeghers syndrome	175200	20301443
N5	Lynch syndrome	120435	20301390

Table continues below

	Typical_age_of_onset	Gene	MIM_gene	Inheritance	Variants_to_report
N1	Adult	BRCA1	113705	AD	KP&EP
N2	Adult	BRCA2	600185	AD	KP&EP
N3	Child/Adult	TP53	191170	AD	KP&EP
N4	Child/Adult	STK11	602216	AD	KP&EP
N5	Adult	MLH1	120436	AD	KP&EP

```
## ACMG-59 Genes:
```

```
##  [1] BRCA1   BRCA2   TP53    STK11   MLH1    MSH2    MSH6    PMS2
##  [9] APC     MUTYH   BMPR1A  SMAD4   VHL     MEN1    RET     PTEN
## [17] RB1     SDHD    SDHAF2  SDHC    SDHB    TSC1    TSC2    WT1
## [25] NF2     COL3A1  FBN1    TGFBR1  TGFBR2  SMAD3   ACTA2   MYH11
## [33] MYBPC3  MYH7    TNNT2   TNNI3   TPM1    MYL3    ACTC1   PRKAG2
## [41] GLA     MYL2    LMNA    RYR2    PKP2    DSP     DSC2    TMEM43
## [49] DSG2    KCNQ1   KCNH2   SCN5A   LDLR    APOB    PCSK9   ATP7B
## [57] OTC     RYR1    CACNA1S
```

1.2 Download ClinVar VCF

ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh37/clinvar.vcf.gz
ClinVar is the central repository for variant interpretations. Relevant information from the VCF includes:
(a) CLNSIG = “Variant Clinical Significance, 0 - Uncertain, 1 - Not provided, 2 - Benign, 3 - Likely benign, 4 - Likely pathogenic, 5 - Pathogenic, 6 - Drug response, 7 - Histocompatibility, 255 - Other”
(b) CLNDBN = “Variant disease name”
(c) CLNDSDBID = “Variant disease database ID”
(d) CLNREVSTAT = “Review Status, no_assertion, no_criteria, single - criterion provided single submitter, mult - criteria provided multiple submitters no conflicts, conf - criteria provided conflicting interpretations, exp - Reviewed by expert panel, guideline - Practice guideline”
(e) INTERP = Pathogenicity (likely pathogenic or pathogenic; CLNSIG = 4 or 5)

```
##                   VAR_ID CHROM       POS ID REF  ALT
## 59912    4_68619662_A_G     4  68619662  .   A    G
## 86327       7_977079_G_A     7    977079  .   G    A
## 122198 10_103530370_C_T    10 103530370  .   C    T
## 122206 10_103534589_G_C    10 103534589  .   G    C
## 189946   17_7579368_A_G    17   7579368  .   A    G
## 232565     X_585258_C_A     X    585258  .   C    A
## 232567     X_585263_G_C     X    585263  .   G    C
## 232569     X_591261_G_A     X    591261  .   G    A
## 232571     X_591695_C_T     X    591695  .   C    T
## 232573     X_591752_G_A     X    591752  .   G    A
## 232575     X_591926_T_G     X    591926  .   T    G
## 232577     X_595354_G_T     X    595354  .   G    T
## 232579     X_595379_G_T     X    595379  .   G    T
## 232581     X_595422_A_G     X    595422  .   A    G
## 232585     X_601571_C_T     X    601571  .   C    T
## 232587     X_601577_G_C     X    601577  .   G    C
## 232589     X_601578_C_A     X    601578  .   C    A
## 232592     X_601772_C_T     X    601772  .   C    T
## 232595     X_605321_G_A     X    605321  .   G    A
## 232598   X_605654_A_AAG     X    605654  .   A  AAG
## 232600    X_1407492_T_C     X   1407492  .   T    C
## 232602    X_1409305_G_C     X   1409305  .   G    C
## 232604    X_1428421_G_T     X   1428421  .   G    T
## 232700    X_8700077_T_C     X   8700077  .   T    C
## 243456  Y_535123_N_NTGT     Y    535123  .   N NTGT
##                                        CLNSIG INTERP
## 59912                              Pathogenic   TRUE
## 86327                              protective  FALSE
## 122198                      Likely_pathogenic   TRUE
## 122206                      Likely_pathogenic   TRUE
## 189946 Likely_benign, Uncertain_significance  FALSE
## 232565                             Pathogenic   TRUE
## 232567                             Pathogenic   TRUE
## 232569                 Uncertain_significance  FALSE
## 232571   Benign, Likely_benign, not_provided  FALSE
## 232573                          Likely_benign  FALSE
## 232575                                 Benign  FALSE
## 232577                          Likely_benign  FALSE
## 232579                             Pathogenic   TRUE
## 232581                      Likely_pathogenic   TRUE
## 232585                              Pathogenic   TRUE
## 232587                              Pathogenic   TRUE
## 232589                              Pathogenic   TRUE
## 232592                              Pathogenic   TRUE
## 232595                 Uncertain_significance  FALSE
## 232598                 Uncertain_significance  FALSE
## 232600                                  Benign  FALSE
## 232602                                  Benign  FALSE
## 232604                                  Benign  FALSE
## 232700                              Pathogenic   TRUE
## 243456                              Pathogenic   TRUE
```

```
## Processed ClinVar data frame 224657 x 18 (selected rows/columns):
```

Table continues below

VAR_ID	CHROM	POS	ID	REF	ALT	CLNSIG	INTERP	GOLD_STARS	pathogenic	benign	CLNREVSTAT	CLNDSDBID
1_955619_G_C	1	955619	.	G	C	Likely_benign	FALSE	1	FALSE	TRUE	1	1
1_957568_A_G	1	957568	.	A	G	Uncertain_significance	FALSE	1	FALSE	TRUE	1	1
1_957605_G_A	1	957605	.	G	A	Likely_benign	TRUE	0	TRUE	FALSE	1	1
1_957640_C_T	1	957640	.	C	T	Uncertain_significance	FALSE	1	FALSE	TRUE	1	1

1.3 Download 1000 Genomes VCFs

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.[chrom].phase3_[version].20130502.genotypes.vcf.gz Downloaded 1000 Genomes VCFs are saved in: /Users/jamesdiao/Documents/Kohane_Lab/2017-ACMG-penetrance/1000G/

```
## Download report: region and successes: 59 x 6 (selected rows):
```

gene	name	chrom	start	end	downloaded
BRCA1	NM_007294	17	41196311	41277500	TRUE
BRCA2	NM_000059	13	32889616	32973809	TRUE
TP53	NM_000546	17	7571719	7590868	TRUE
STK11	NM_000455	19	1205797	1228434	TRUE
MLH1	NM_000249	3	37034840	37092337	TRUE

```
## File saved as download_output.txt in Supplementary_Files
```

1.4 Import and Process 1000 Genomes VCFs

(a) Unnest the data frames to 1 row per variant_ID key (CHROM_POSITION_REF_ALT).
(b) Remove all insertions, deletions, CNV, etc, and keep only missense variants (1 REF, 1 ALT)
(c) For 1000 Genomes: convert genomes to allele counts. For example: (0|1) becomes 1, (1|1) becomes 2. Multiple alleles are unnested into multiple counts. For example: (0|2) becomes 0 for the first allele (no 1s) and 1 for the second allele (one 2).

```
## Processed 1000 Genomes VCFs: 141467 x 2516 (selected rows/columns):
```

GENE	AF_1000G	VAR_ID	CHROM	POS	ID	REF	ALT
BRCA1	0.004193290	17_41196363_C_T	17	41196363	rs8176320	C	T
BRCA1	0.008386580	17_41196368_C_T	17	41196368	rs184237074	C	T
BRCA1	0.000998403	17_41196372_T_C	17	41196372	rs189382442	T	C
BRCA1	0.342252000	17_41196408_G_A	17	41196408	rs12516	G	A
BRCA1	0.000399361	17_41196409_G_C	17	41196409	rs548275991	G	C

Table continues below

HG00096	HG00097	HG00099	HG00100	HG00101	HG00102
0	0	0	0	0	0
0	0	0	0	0	0
0	0	0	0	0	0
1	0	1	1	0	2
0	0	0	0	0	0

1.5 Import and Process gnomAD/ExAC VCFs

(a) Unnest the data frames to 1 row per variant_ID key (CHROM_POSITION_REF_ALT).
(b) Remove all insertions, deletions, CNV, etc, and keep only missense variants (1 REF, 1 ALT)
(c) Collect superpopulation-level allele frequencies: African = AFR, Latino = AMR, European (Finnish + Non-Finnish) = EUR, East.Asian = EAS, South.Asian = SAS.

```
## Processed gnomAD VCFs: 96742 x 48 (selected rows/columns):
```

	GENE	AF_GNOMAD	VAR_ID
210313	CACNA1S	0.00003170	1_201047100_A_G
3537	BRCA2	0.00001290	13_32914977_A_G
24258	RB1	0.00000397	13_49037941_A_C
70823	KCNQ1	0.00000868	11_2482624_A_G
53404	TPM1	0.00729000	15_63362179_G_A

```
## Processed ExAC VCFs: 59883 x 45 (selected rows/columns):
```

	GENE	AF_EXAC	VAR_ID
22159	NF2	0.000031180	22_30069425_G_A
31792	MYH7	0.000024740	14_23898940_C_T
269110	APOB	0.000041890	2_21260121_C_T
161004	ATP7B	0.000008299	13_52508899_T_G
71415	CACNA1S	0.000016470	1_201029795_G_A

1.6 Collect 1000 Genomes Phase 3 Populations Map

This will allow us to assign genotypes from the 1000 Genomes VCF to ancestral groups.
From: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/integrated_call_samples_v3.20130502.ALL.panel

```
## Phase 3 Populations Map Table: 2504 x 4 (selected rows)
```

sample	pop	super_pop	gender
HG02702	GWD	AFR	male
NA19309	LWK	AFR	male
HG00530	CHS	EAS	male
NA20785	TSI	EUR	male
NA12046	CEU	EUR	female
HG04219	ITU	SAS	male

1.7 Merge ClinVar with gnomAD, ExAC, and 1000 Genomes

```
## Breakdown of ClinVar Variants
```

Subset_ClinVar	Number_of_Variants
Total ClinVar	224657
LP/P	43321
ACMG LP/P	9229
ACMG LP/P in gnomAD	437
ACMG LP/P in ExAC	283
ACMG LP/P in 1000 Genomes	27

```
## Breakdown of ACMG-gnomAD Variants
```

Subset_gnomAD	Number_of_Variants
ACMG in gnomAD	96742
ClinVar-ACMG in gnomAD	14517
LP/P-ACMG in gnomAD	437

```
## Breakdown of ACMG-ExAC Variants
```

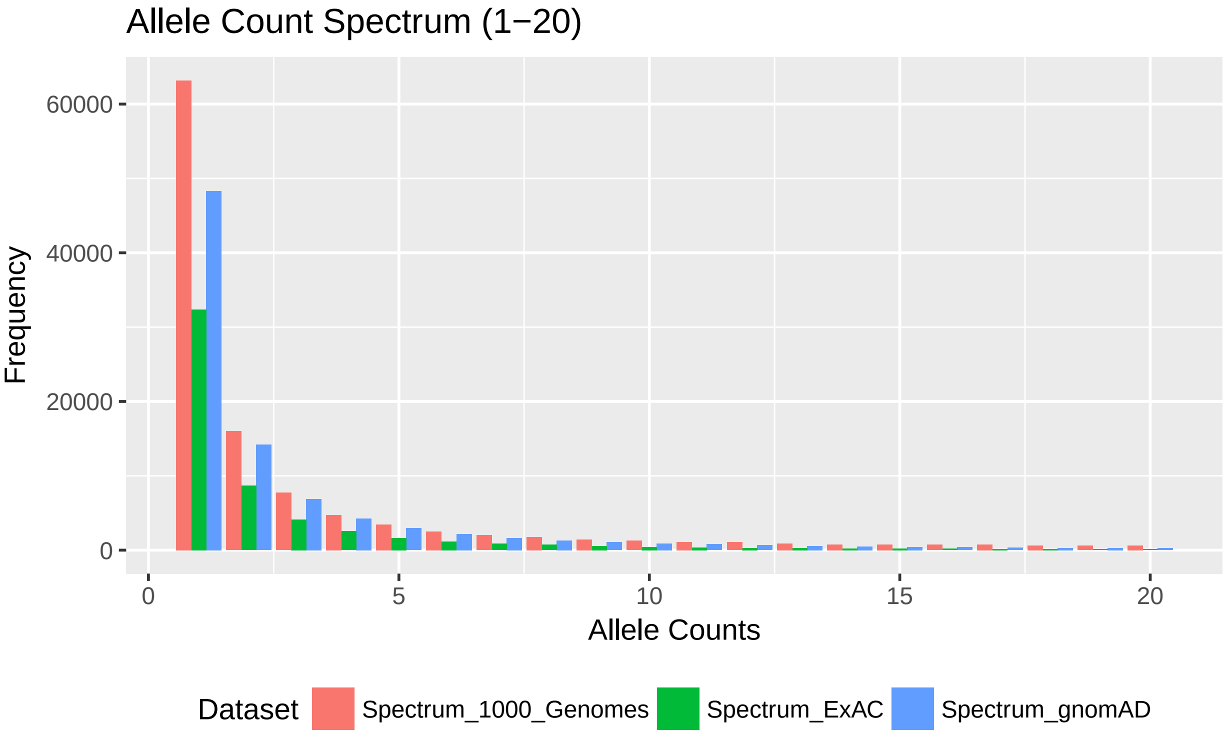
Subset_gnomAD	Number_of_Variants
ACMG in ExAC	59883
ClinVar-ACMG in ExAC	11155
LP/P-ACMG in ExAC	283

```
## Breakdown of ACMG-1000G Variants
```

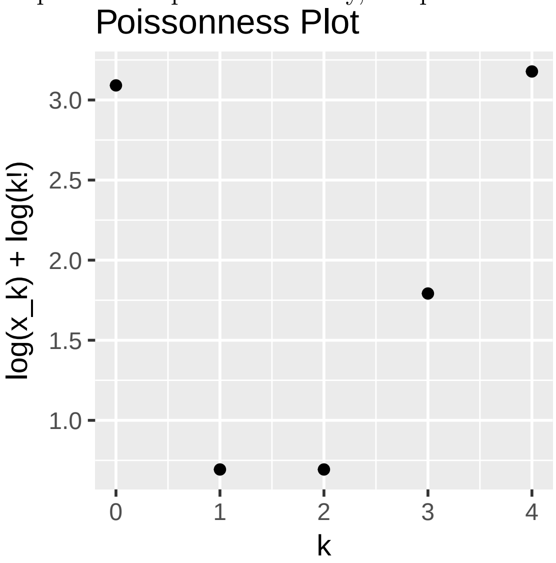
Subset_gnomAD	Number_of_Variants
ACMG in 1000G	141466
ClinVar-ACMG in 1000G	6080
LP/P-ACMG in 1000G	27

2 Plot Summary Statistics Across Populations

2.1 Distribution of Allele Counts



We can model this as a Poisson binomial- the summed occurance of variants with different allele frequencies. If we assume that the allele frequencies are approximately the same and that variants are independent, (may not be good assumptions), then the distribution follows Binom(n,p), n = # samples and p = allele frequency. Because n is large and p is small, we can then use a Poisson approximation to the binomial.
The fit of this approximation may be tested by the Poissonness plot (Hoaglin 1980), or $log(x_k) + log(k!)$ vs. k.
If $x_k = n\,\Pr(X = k) = n\left(\frac{\lambda^k e^{-k}}{k!}\right)$, then $\ln x_k + \ln k! = \ln n + k \ln \lambda - \lambda =$ linear function of k.
Despite some upward concavity, the plot demonstrates reasonable Poissonness, with correlation = 0.16.



2.2 Overall Non-Reference Sites

2.2.0.1 For 1000 Genomes

Each individual has n non-reference sites, which can be found by counting. The mean number is computed for each population.

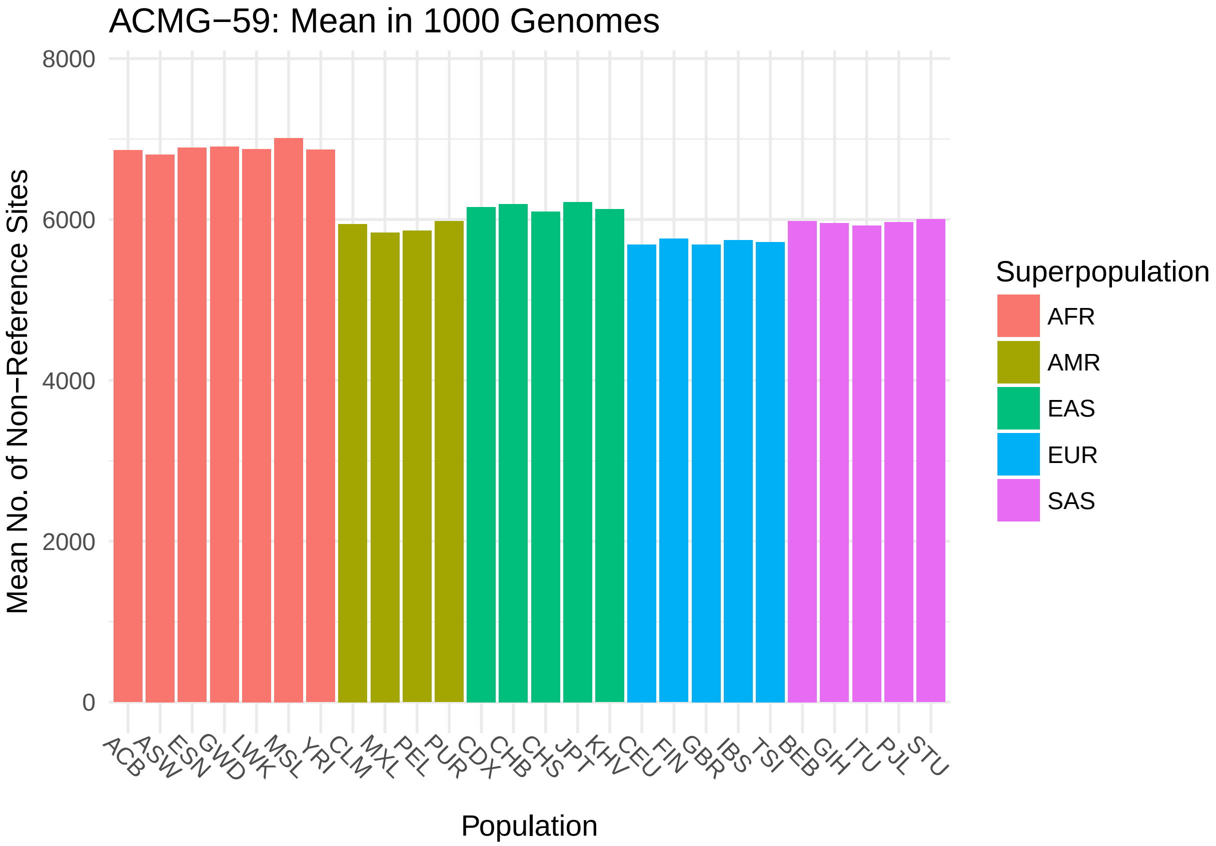
Ex: the genotype of 3 variants in 3 people looks like this:

	HG00366	HG00367	HG00368
Variant 1	2	1	1
Variant 2	2	1	1
Variant 3	1	0	0

Count the number of non-reference sites per individual:

HG00366	HG00367	HG00368
3	2	2

```
## Mean = 2.33
```



Note: the error bars denote standard deviation, not standard error.

2.2.0.2 For gnomAD/ExAC

The mean number of non-reference sites is $E(V)$, where $V = \sum_{i=1}^{n} v_i$ is the number of non-reference sites at all variant positions v_1 through v_n.

At each variant site, the probability of having at least 1 non-reference allele is $P(v_i) = P(v_{i,a} \cup v_{i,b})$, where a and b indicate the 1st and 2nd allele at each site.

If the two alleles are independent, $P(v_{i,a} \cup v_{i,b}) = 1 - (1 - P(v_{i,a}))(1 - P(v_{i,b})) = 1 - (1 - AF(v_i))^2$

If all variants are independent, $E(V) = \sum_{i=1}^{n} 1 - (1 - AF(v_i))^2$ for any set of allele frequencies.

Ex: the allele frequencies of 3 variants across the 5 superpopulations looks like this:

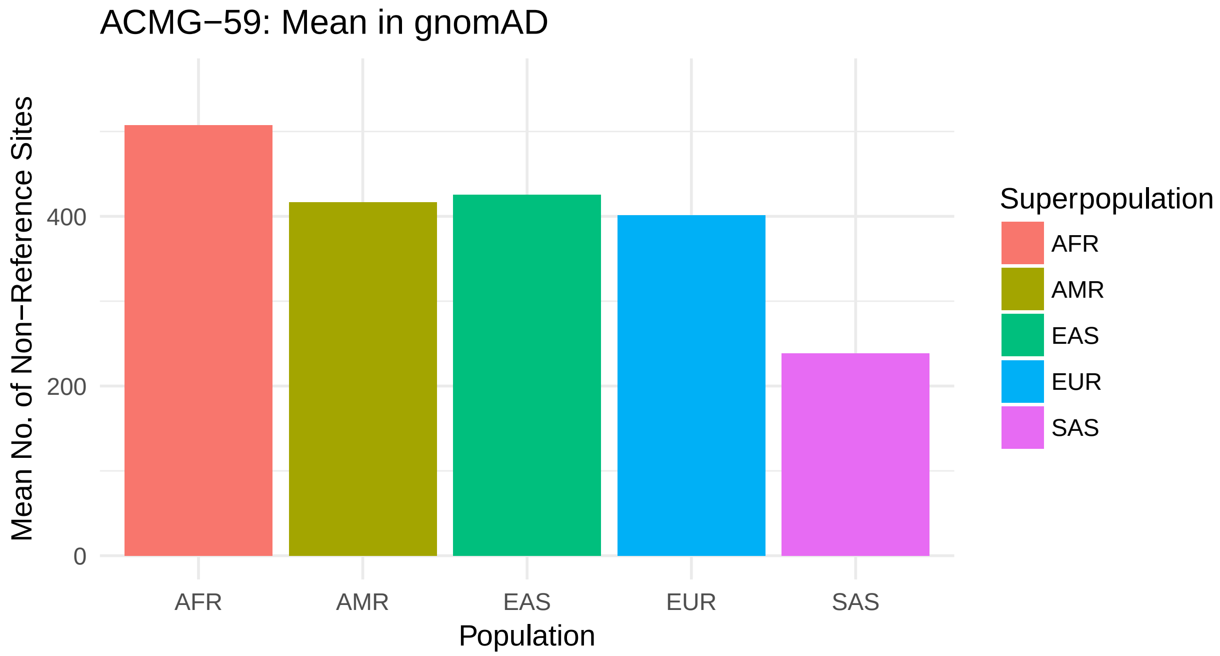
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

The probability of having at least 1 non-reference site at each variant - (0|1) (1|0) or (1|1) is given by $1 - (1 - AF)^2$. Note that this is approximately $2 * AF$ when AF is small:

	AFR	AMR	EAS	EUR	SAS
Variant 1	0.19	0.36	0	0	0.51
Variant 2	0.36	0	0.51	0	0.19

By linearity of expectation, the expected (mean) number of non-reference sites is $\sum E(V_i) = \sum(columns)$.

AFR	AMR	EAS	EUR	SAS
0.55	0.36	0.51	0	0.7



2.3 Fraction of Individuals with Pathogenic Sites

2.3.0.1 For 1000 Genomes

We can count up the fraction of individuals with 1+ non-reference site(s) in each population. This is the fraction of individuals who would receive a positive genetic test result in at least 1 of the ACMG-59 genes.

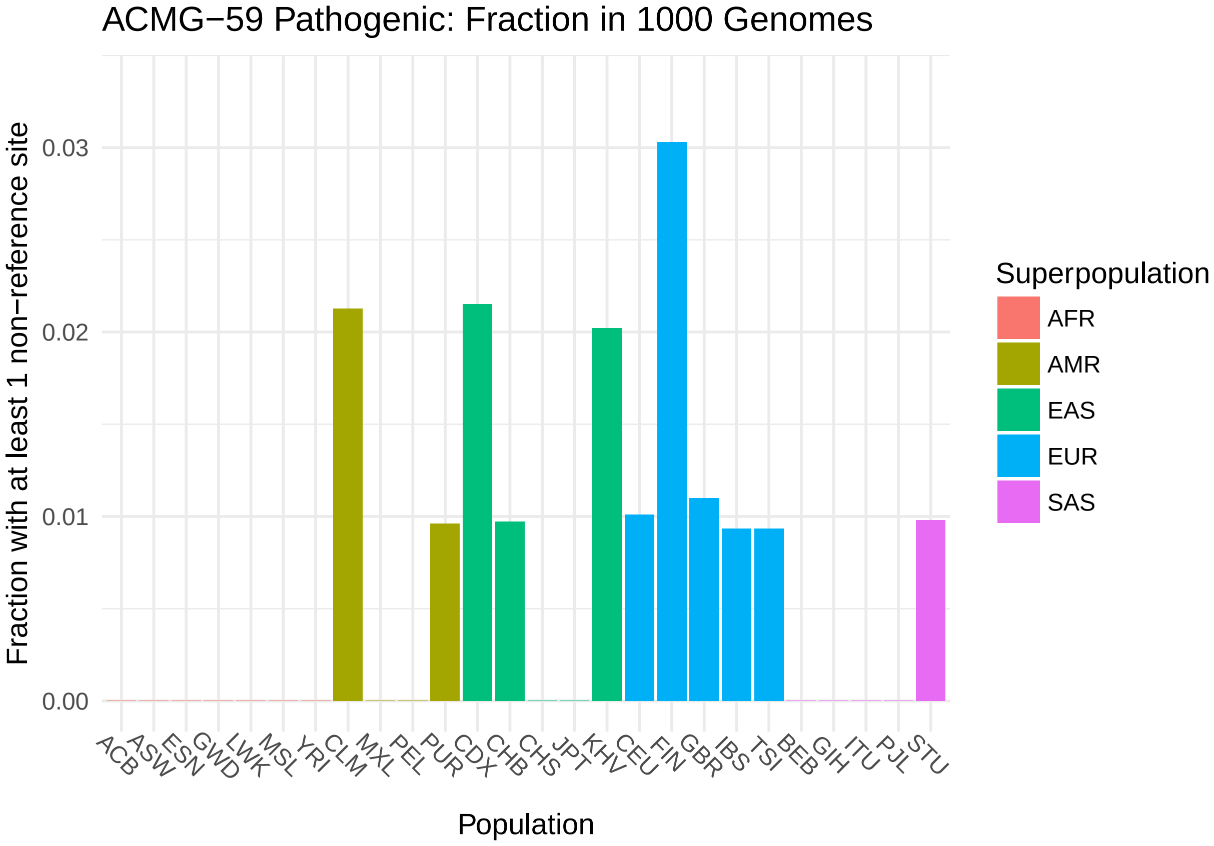
Ex: the genotype of 3 variants in 3 people looks like this:

	HG00366	HG00367	HG00368
Variant 1	2	1	1
Variant 2	2	1	1
Variant 3	1	0	0

Count each individual as having a non-reference site (1) or having only reference sites (0):

HG00366	HG00367	HG00368
1	1	1

```
## Mean = 1
```



2.3.0.2 For gnomAD/ExAC

The probability of having at least 1 non-reference site is $P(X)$, where X indicates a non-reference site at any variant position v_1 through v_n.

Recall that $P(v_i) = P(v_{i,a} \cup v_{i,b}) = 1 - (1 - AF(v))^2$ when alleles are independent.

If all alleles are independent, $P(X) = P(\bigcup_{i=1}^{n} v_i) = 1 - \prod_{i=1}^{n}(1 - AF(v_i))^2$

Ex: the allele frequencies of 3 variants across the 5 superpopulations looks like this:

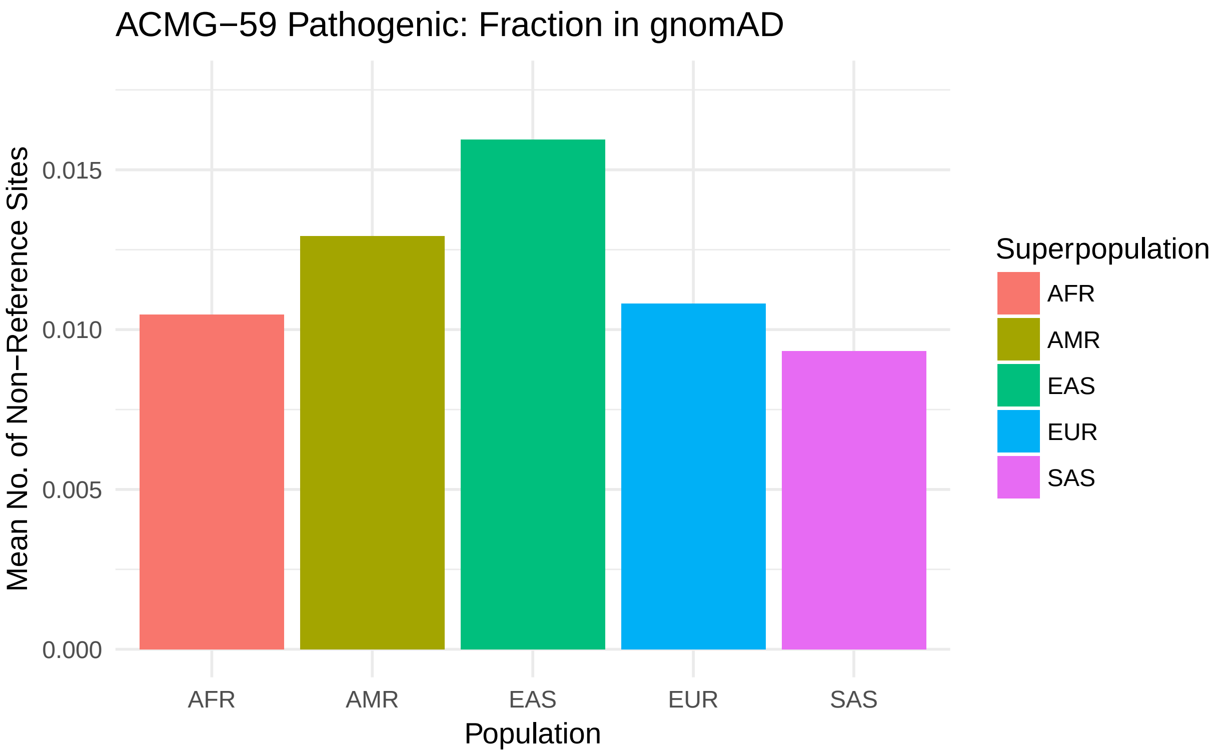
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

The probability of having at least 1 non-reference site at each variant - (0|1) (1|0) or (1|1) is given by $1 - (1 - AF)^2$. Note that this is approximately $2 * AF$ when AF is small:

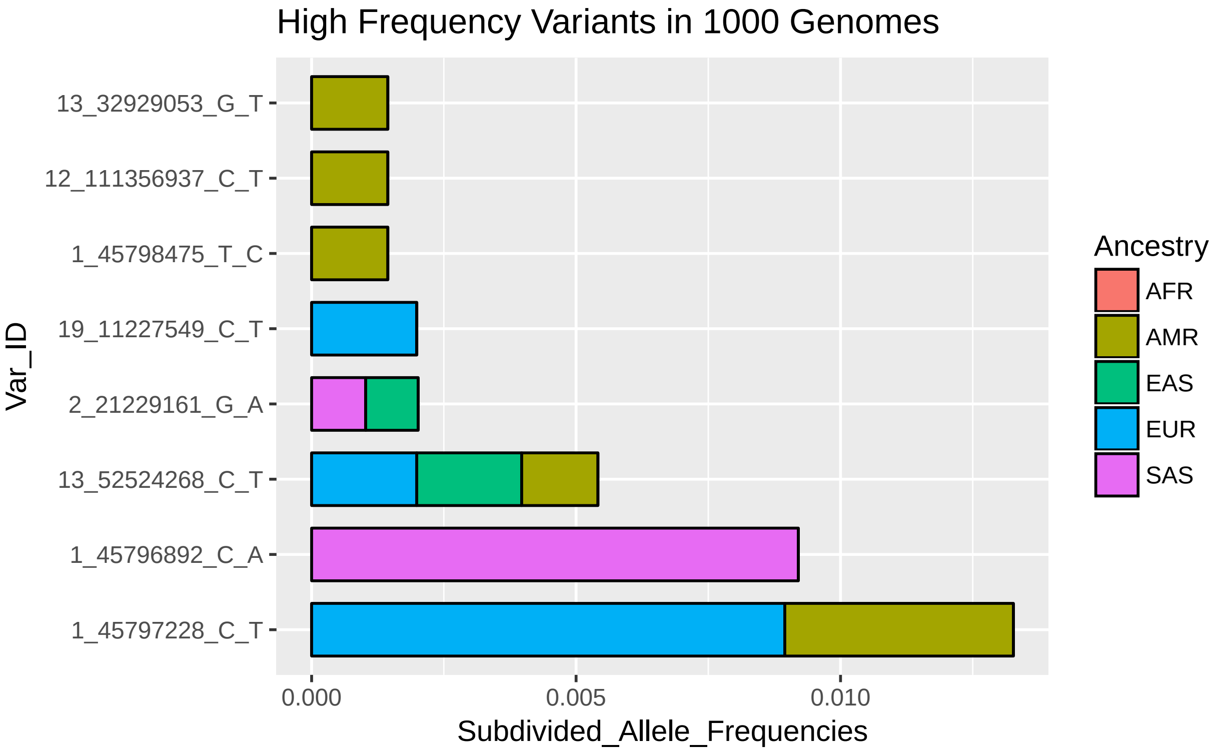
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.19	0.36	0	0	0.51
Variant 2	0.36	0	0.51	0	0.19

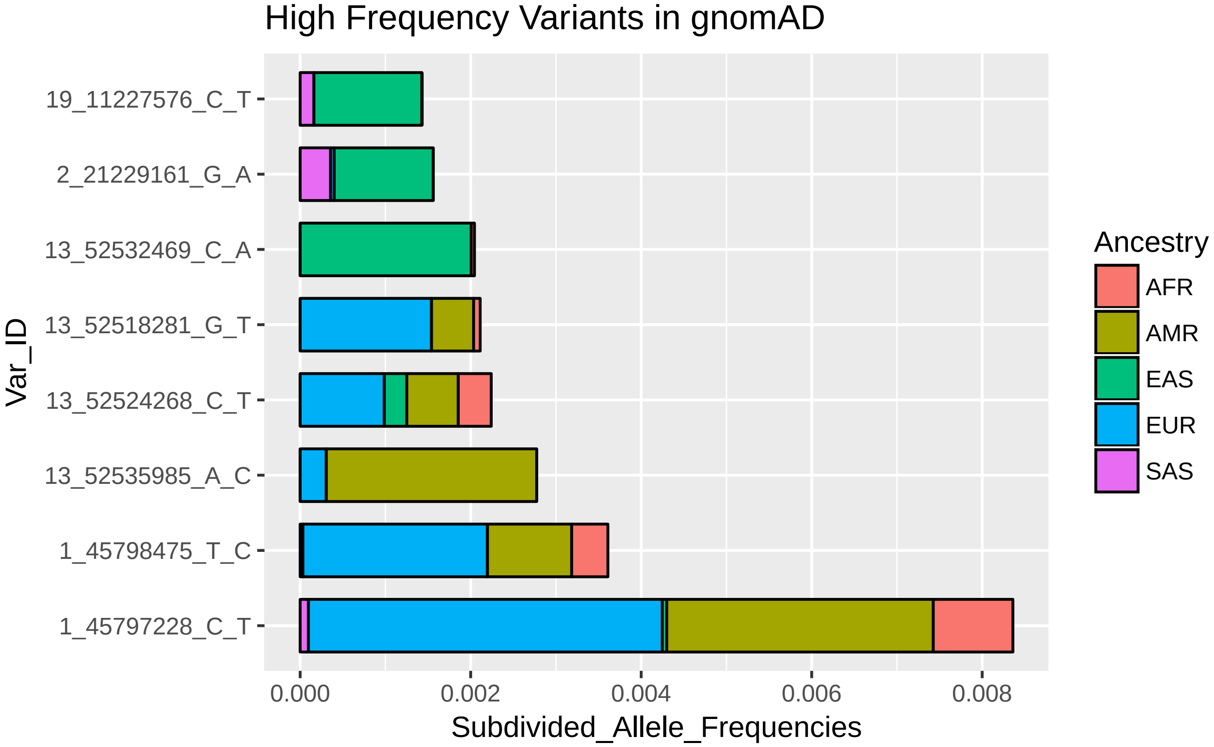
The expected (mean) number of non-reference sites is given by $1 - \prod(1 - AF)^2$.

AFR	AMR	EAS	EUR	SAS
0.4816	0.36	0.51	0	0.6031



2.4 Common Pathogenic Variants by Ancestry





3 Penetrance Estimates

3.1 Bayes' Rule as a Model for Estimating Penetrance

Let V_x be the event that an individual has 1 or more variant related to disease x,
and D_x be the event that the individual is later diagnosed with disease x.

In this case, we can define the following probabilities:
1. Prevalence = $P(D_x)$
2. Population Allele Frequency (PAF) = $P(V_x)$
3. Case Allele Frequency (CAF) = $P(V_x|D_x)$
4. Penetrance = $P(D_x|V_x)$

By Bayes' Rule, the penetrance of a variant related to disease x may be defined as:

$$P(D_x|V_x) = \frac{P(D_x) * P(V_x|D_x)}{P(V_x)} = \frac{(Prevalence)(Population\ Allele\ Frequency)}{(Case\ Allele\ Frequency)}$$

To compute penetrance estimates for each of the diseases related to the ACMG-59 genes, we will use the prevalence data we collected into `Literature_Prevalence_Estimates.csv`, allele frequency data from 1000 Genomes/ExAC/gnomAD, and a broad range of values for case allele frequency.

3.2 Import Literature-Based Disease Prevalence Data

Data Collection:
1. Similar disease subtypes were grouped together (e.g., the 8 different types of familial hypertrophic cardiomyopathy), resulting in 30 disease categories across 59 genes.
2. The search query "[disease name] prevalence" was used to find articles using Google Scholar.
3. Prevalence estimates were recorded along with URL, journal, region, publication year, sample size, first author, population subset (if applicable), date accessed, and potential issues. Preference was given to studies with PubMed IDs, more citations, and larger sample sizes.

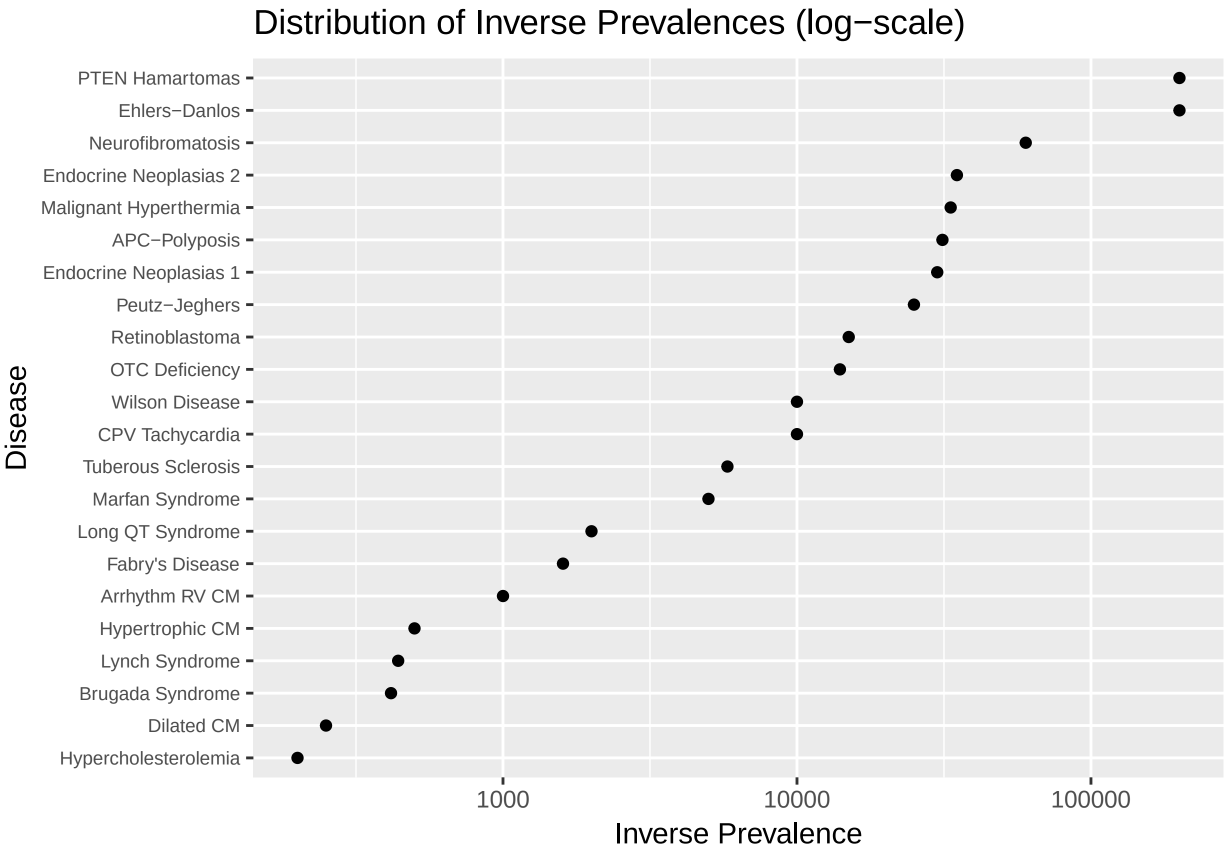
Prevalence was recorded as reported: either a point estimate or a range. Values of varying quality were collected across all diseases.

```
## Table of Literature-Based Estimates 22 x 20 (selected rows/columns):
```

Gene	Phenotype	Inverse_Prevalence	Case_Allele_Frequency
APC	Familial adenomatous polyposis	31250	0.9
MEN1	Multiple endocrine neoplasia type 1	30000	0.9
MYH7|TPM1|MYBPC3|PRKAG2|TNNI3|MYL3|MYL2|ACTC1	Hypertrophic cardiomyopathy	500	0.6
STK11	Peutz-Jeghers syndrome	25000	0.96

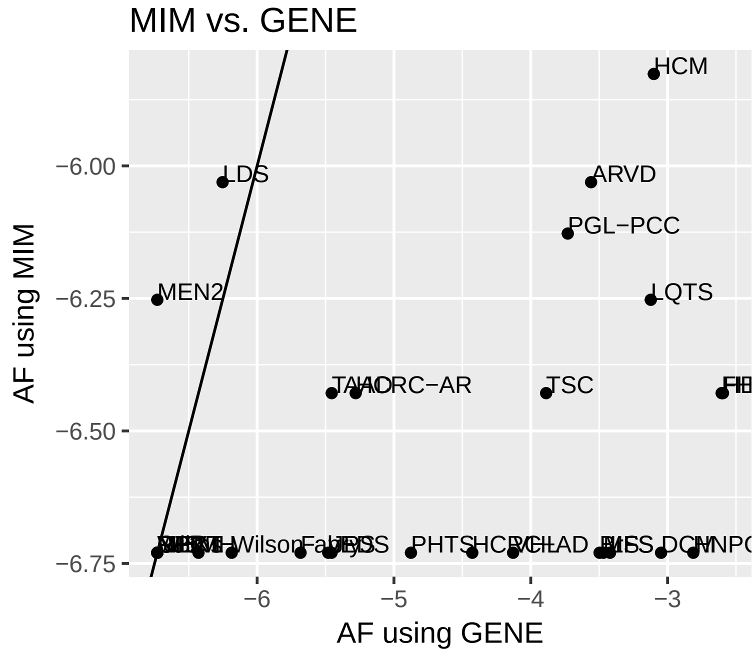
Table continues below

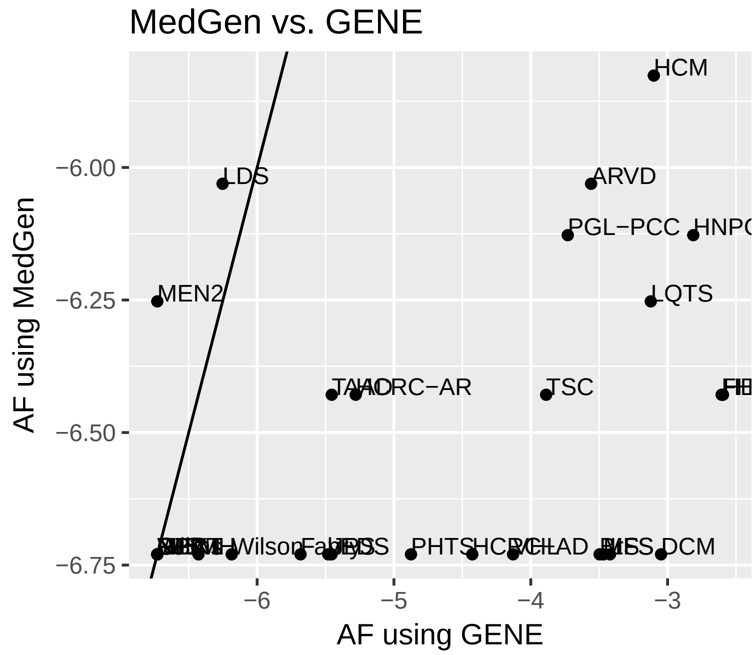
3.3 Distribution of Prevalences



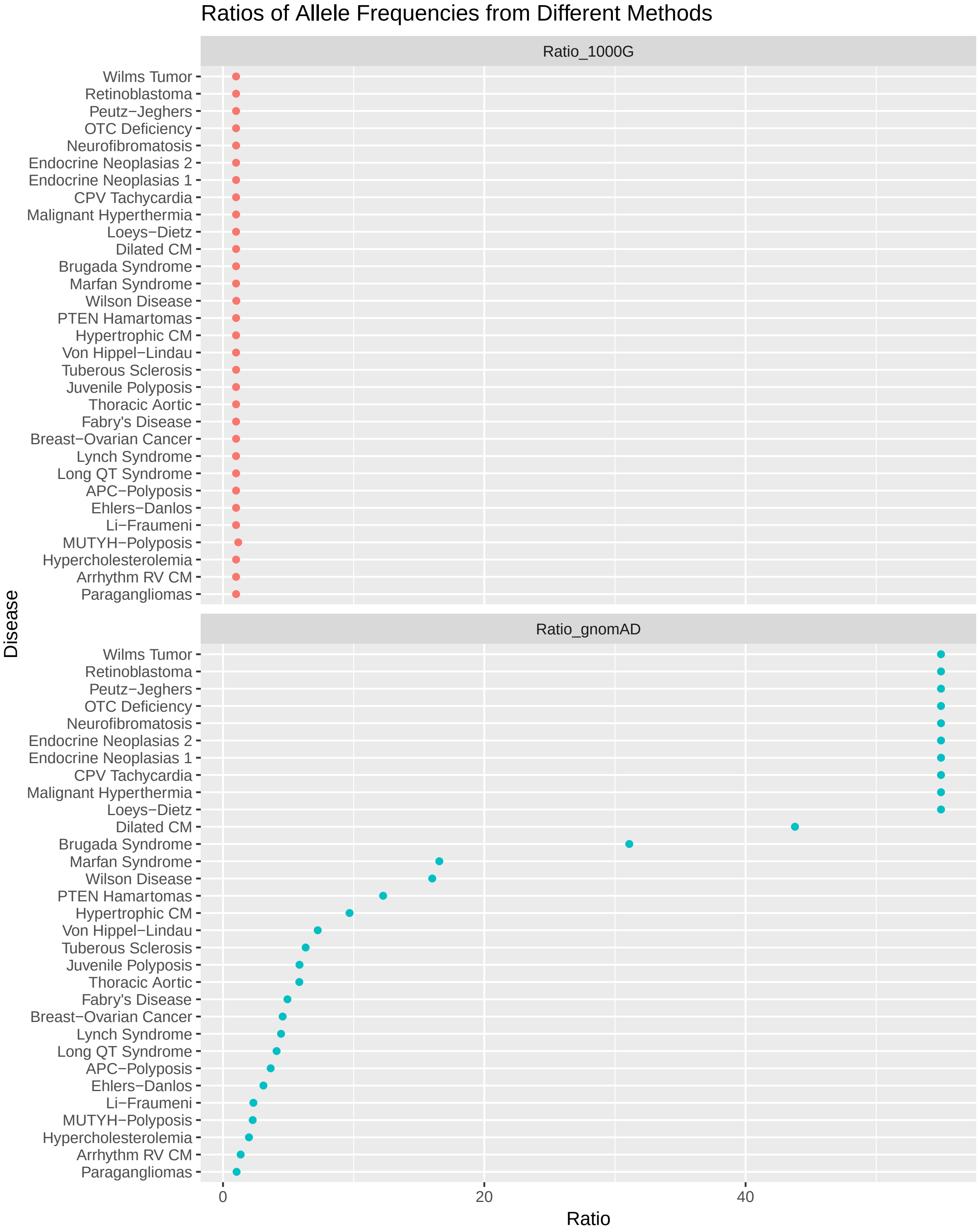
3.4 Collect and Aggregate Allele Frequencies at the Disease-Level

We define AF(disease) as the probability of having at least 1 variant associated with the disease. The variants can be assigned to diseases in two ways:
(1) By associating it by MIM. An MIM code is assigned for around 31% of assertions in each dataset.
(1) By associating it by MedGen. An MIM code is assigned for around 22% of assertions in each dataset.
(2) By associating it by gene. All variants are associated with genes, but some variants may be designated as pathogenic for non-ACMG conditions.
The frequencies across the relevant variants can be aggregated in two ways:
(1) By direct counting, from genotype data in 1000 Genomes.
(2) AF(disease) $= 1 - \prod_{variant}(1 - AF_{variant})$, from population data in 1000 Genomes, ExAC, or gnomAD (assumes independence).

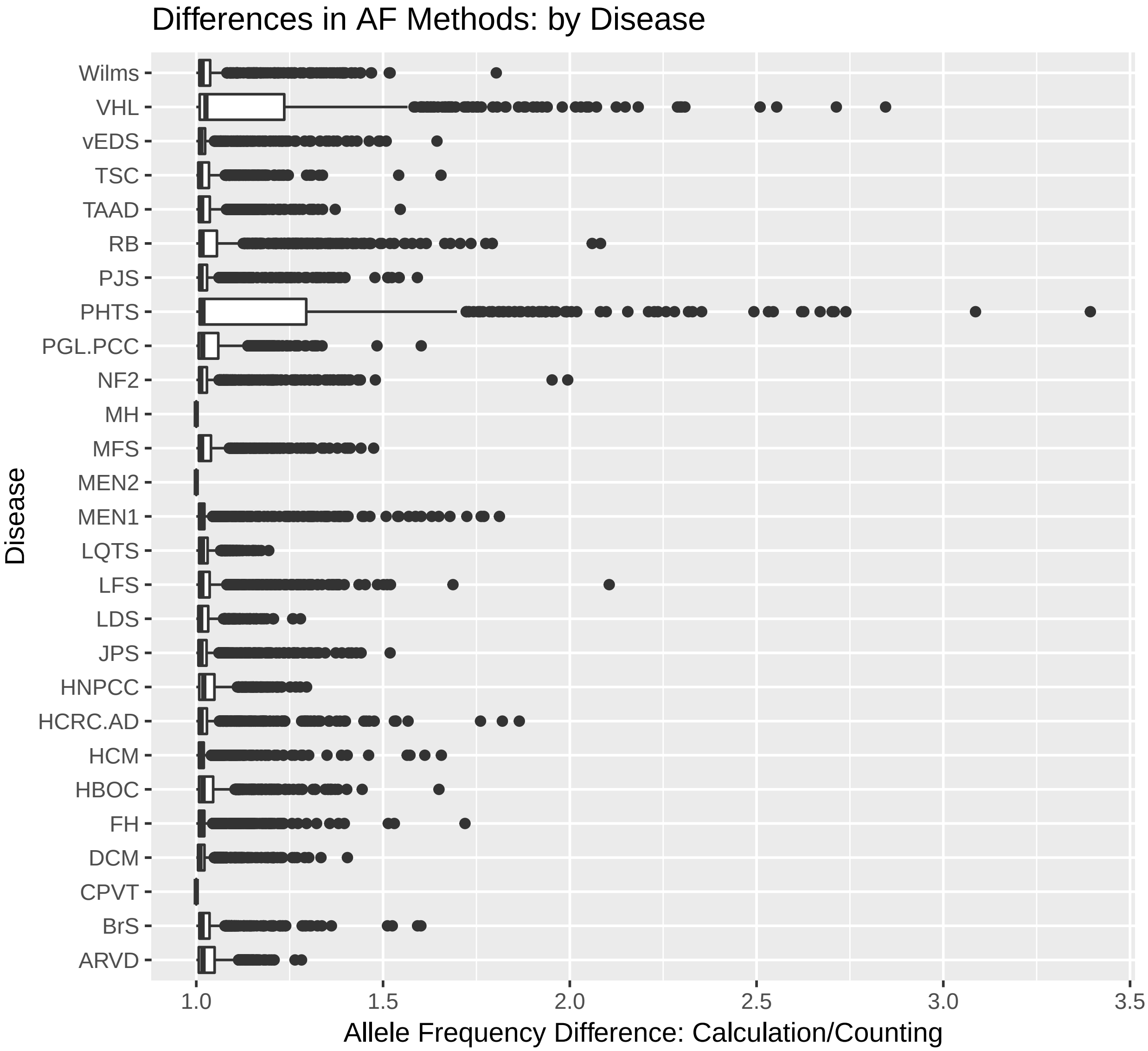


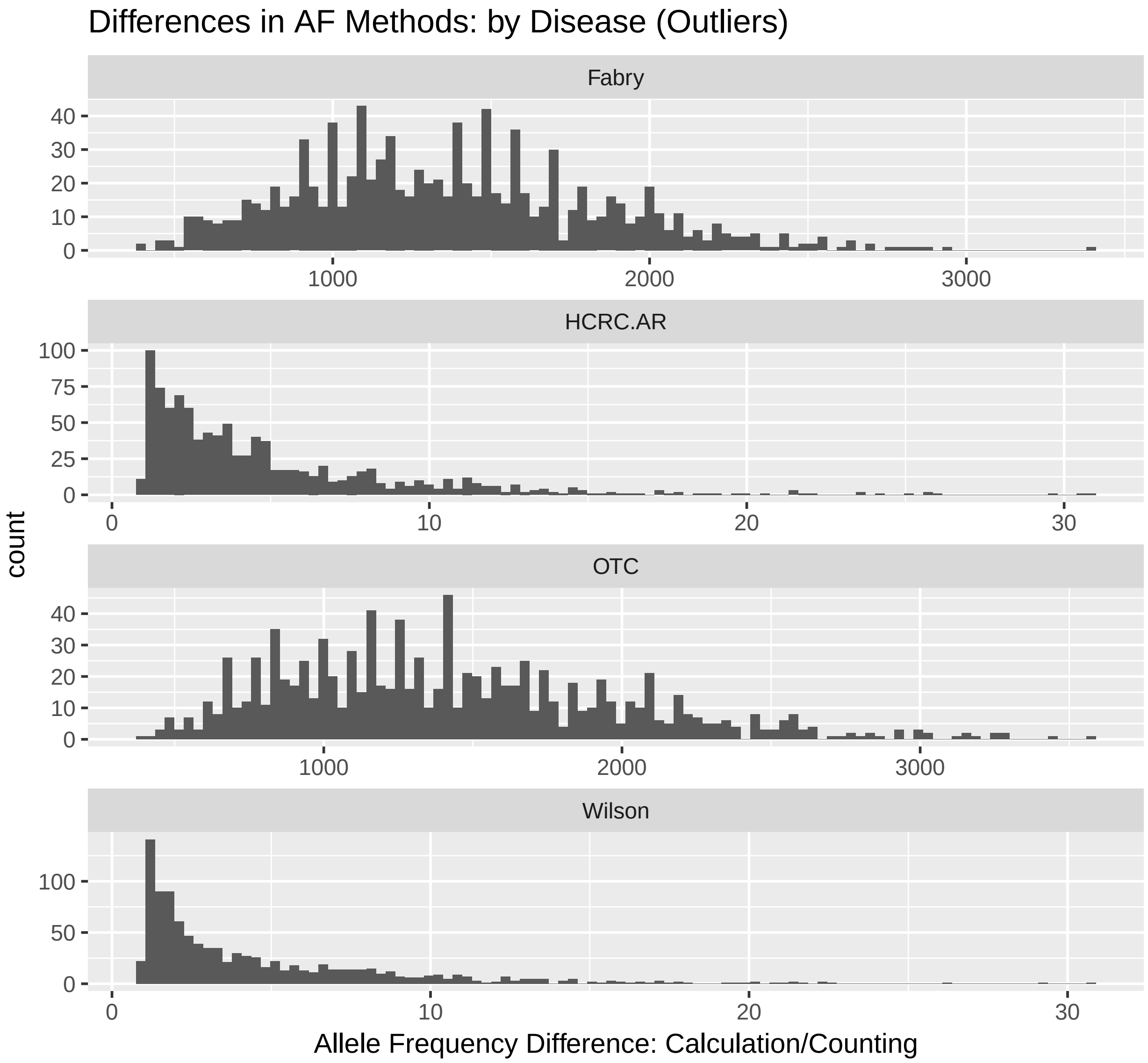


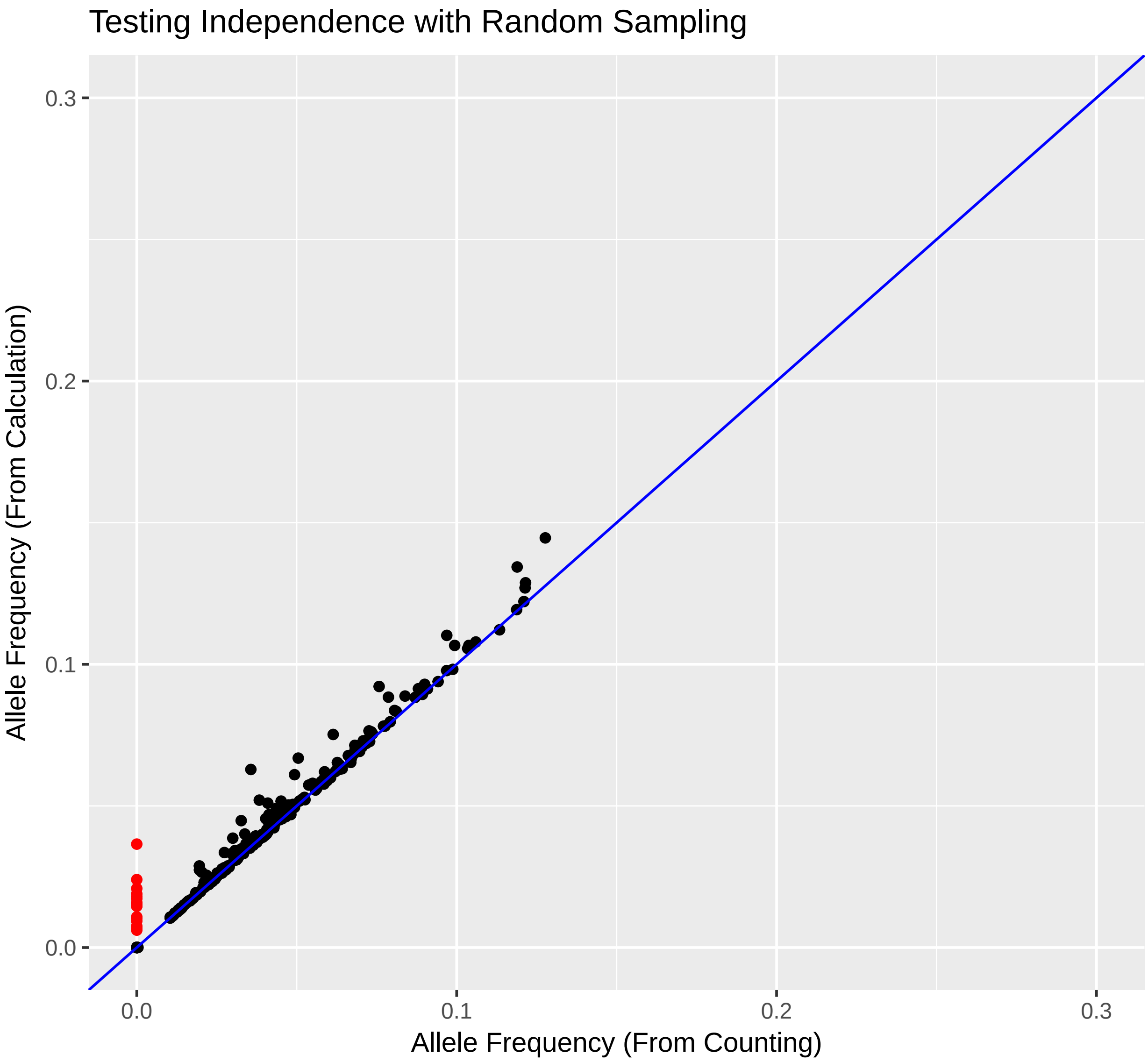
Ratio_1000G (red, top) computes AF(calculation in 1000 Genomes) / AF(counting in 1000 Genomes).
Ratio_gnomAD (blue, bottom) computes AF(calculation in gnomAD) / AF(calculation in 1000 Genomes).



Sampling 1000 variants from all variants in 1000 Genomes to test deviations from independence assumptions. Repeat for 1000 trials and plot the distribution of disease-level allele frequencies (1000 points per disease). Only variants with allele frequency $< 1\%$ are evaluated. Since we look at 17 variants per disease, the maximum is approximately $1 - (1 - 0.01)^{34} \approx 0.29$




Differences in AF Methods: by Disease (Outliers)
Fabry
40
30
20
10
0
1000
2000
3000
HCRC.AR
100
75
50
25
0
0
10
20
30
count
OTC
40
30
20
10
0
1000
2000
3000
Wilson
100
50
0
0
10
20
30
Allele Frequency Difference: Calculation/Counting

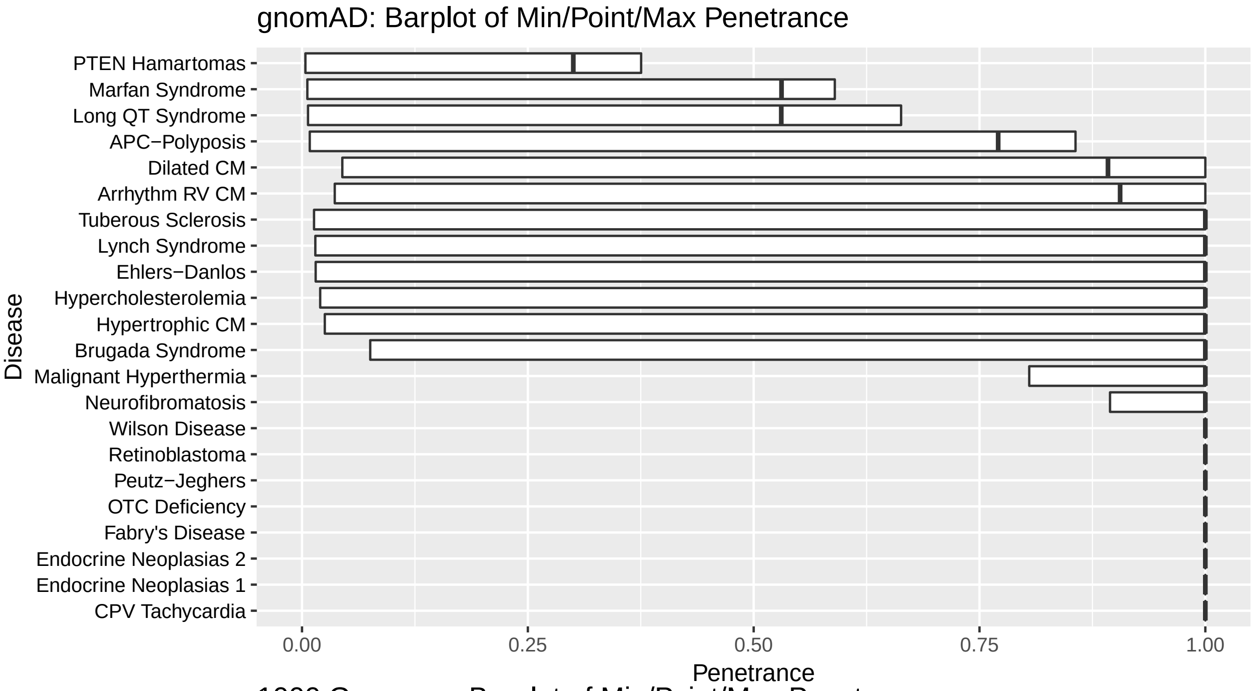


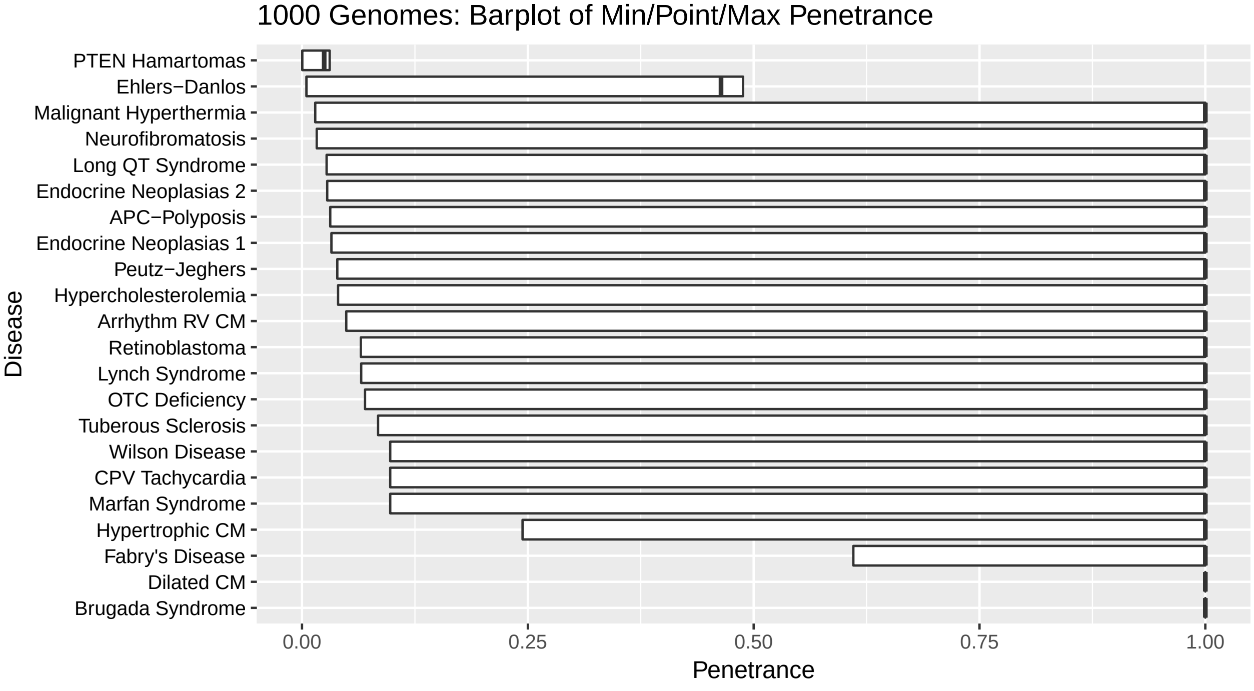
```
## 31 diseases x 1000 points = 31,000 points.
## This plot has been downsampled 100x and contains 310 points.
## AR (autosomal recessive) and XL (X-linked) diseases are colored in red.

## Pearson correlation: 0.989
```

3.5 Penetrance as a Function of P(V|D)

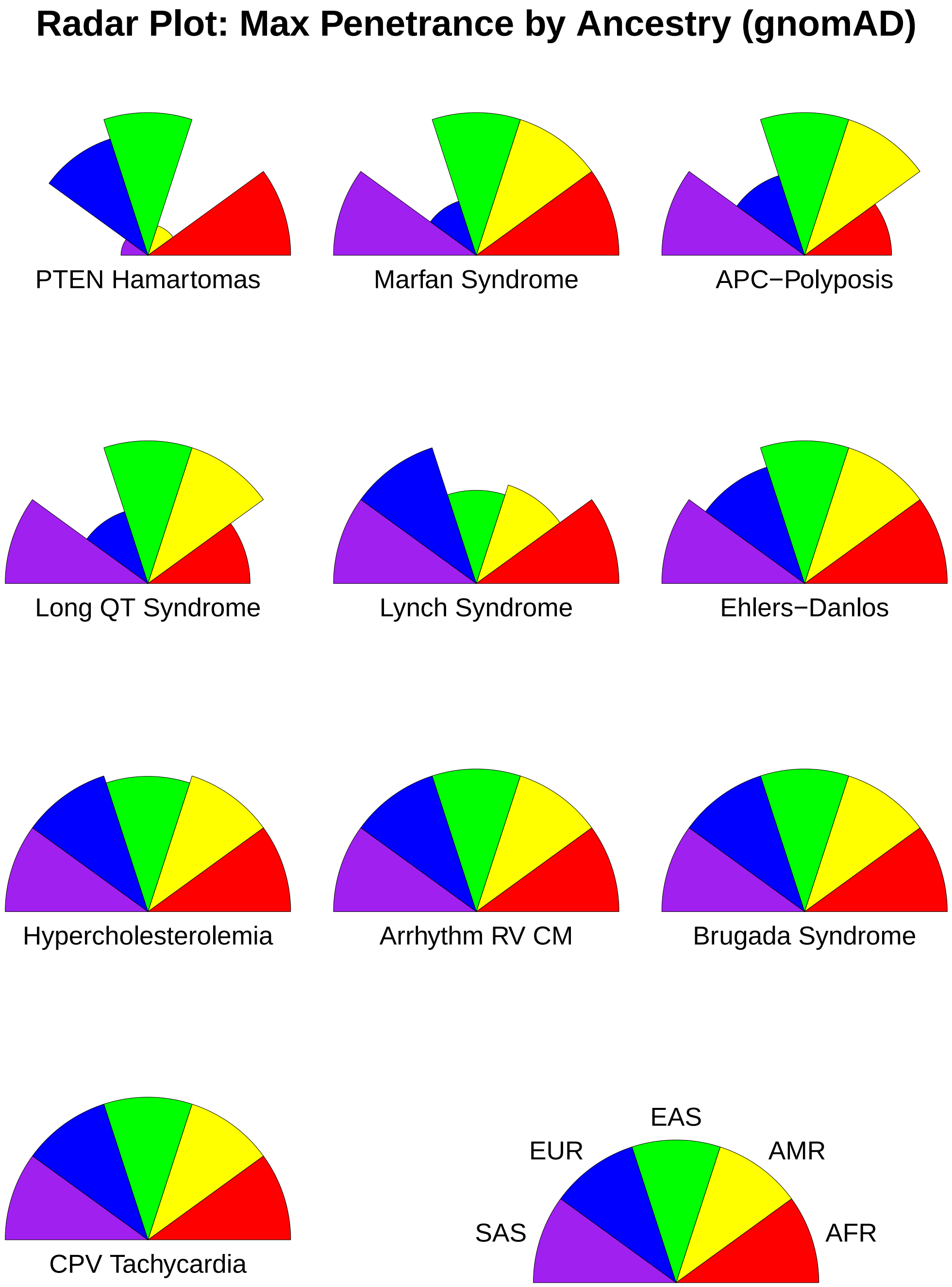
The left end of the boxplot indicates P(V|D) = 0.01,
the bold line in the middle indicates P(V|D) = point value,
the right end of the boxplot indicates P(V|D) = 1.





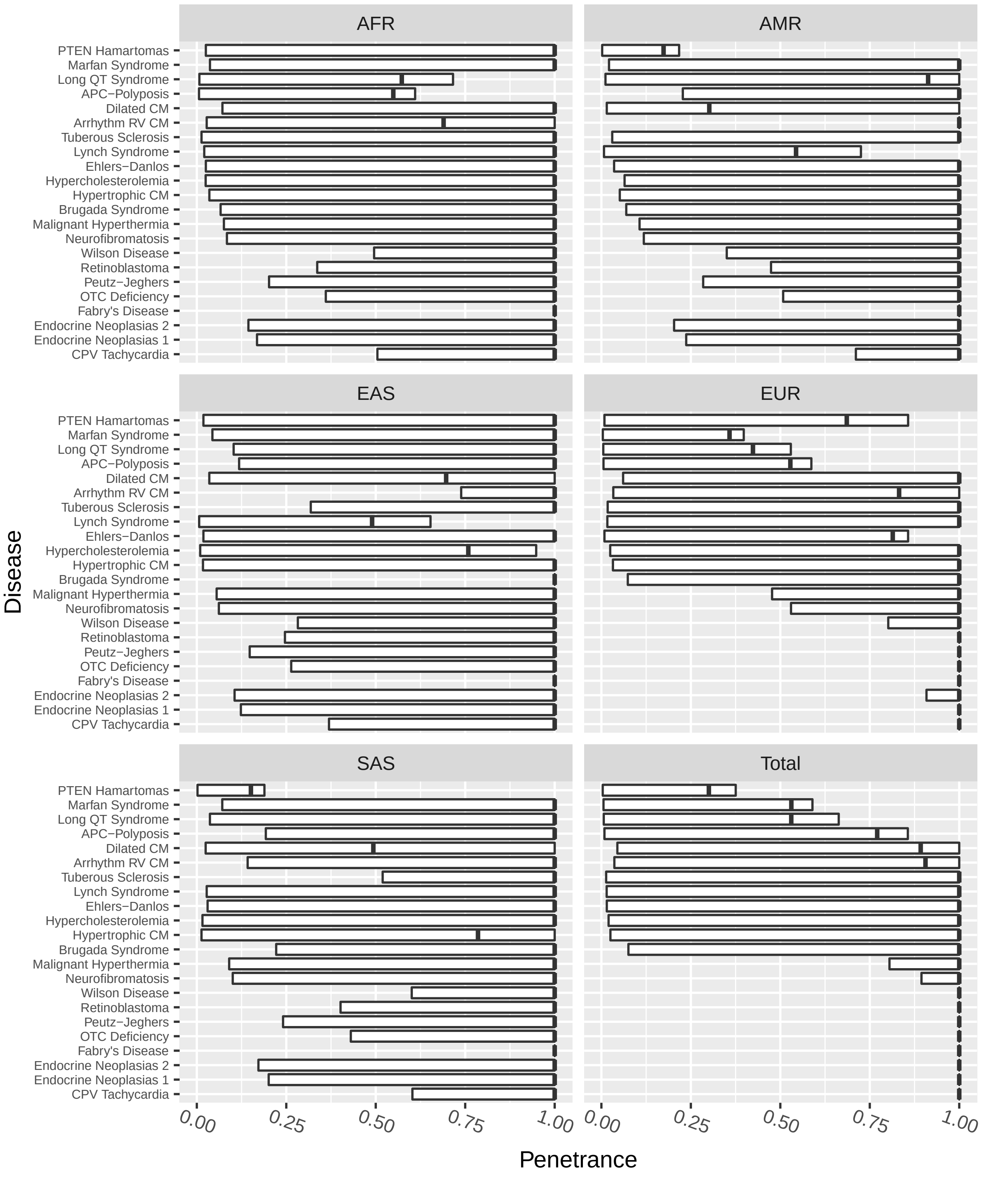
Note: Some diseases have mean theoretical penetrance = 1 because the assumed allelic heterogeneity is greater than is possible, given the observed prevalence and allele frequencies.

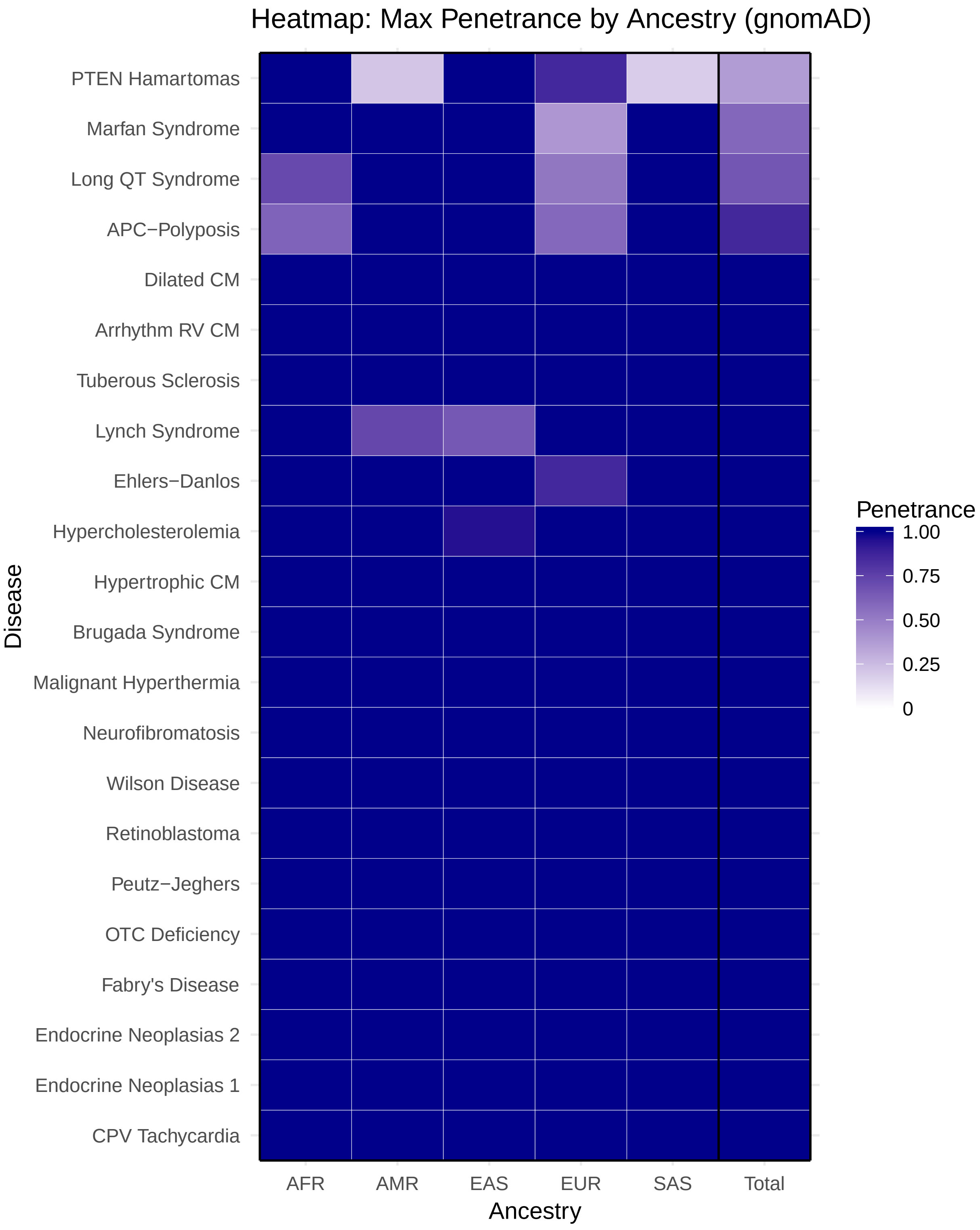
3.6 Penetrance Estimates by Ancestry



```
## [1] These are the top 10 diseases by summed allele frequencies. NULL values are not plotted.
## [1] Each radius is proportional to the penetrance of the disease in the given population.
```

Barplot: Penetrance by Ancestry (gnomAD)




Heatmap: Max Penetrance by Ancestry (gnomAD)
PTEN Hamartomas
Marfan Syndrome
Long QT Syndrome
APC−Polyposis
Dilated CM
Arrhythm RV CM
Tuberous Sclerosis
Lynch Syndrome
Ehlers−Danlos
Hypercholesterolemia
Hypertrophic CM
Brugada Syndrome
Malignant Hyperthermia
Neurofibromatosis
Wilson Disease
Retinoblastoma
Peutz−Jeghers
OTC Deficiency
Fabry's Disease
Endocrine Neoplasias 2
Endocrine Neoplasias 1
CPV Tachycardia
Disease
AFR
AMR
EAS
EUR
SAS
Total
Ancestry
Penetrance
1.00
0.75
0.50
0.25
0