ACMG-ClinVar Penetrance RMarkdown

James Diao, under the supervision of Arjun Manrai

November 22, 2016

Contents

Working Directory: /Users/jamesdiao/Documents/Kohane_Lab/2016-paper-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/

```
## Processed Table from ACMG Website 60 x 8 (selected rows):
```

	Phenotype	Typical_age_of_onset	Gene
N1	Hereditary breast and ovarian cancer	Adult	BRCA1
N2	Hereditary breast and ovarian cancer	Adult	BRCA2
N3	Li-Fraumeni syndrome	Child/Adult	TP53
N4	Peutz-Jeghers syndrome	Child/Adult	STK11
N5	Lynch syndrome	Adult	MLH1

Table continues below

	Inheritance	Variants_to_report
N1	AD	KP&EP
N2	AD	KP&EP
N3	AD	KP&EP
N4	AD	KP&EP
N5	AD	KP&EP

```
## ACMG-56 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) INTERP = Pathogenicity (likely pathogenic or pathogenic; CLNSIG = 4 or 5)

```
## Processed ClinVar data frame 126349 x 14 (selected rows/columns):
```

VAR_ID	CHROM	POS	ID	REF	ALT	CLNSIG
1_949523_C_T	1	949523	rs786201005	C	T	5
1_949739_G_T	1	949739	rs672601312	G	T	5
1_955597_G_T	1	955597	rs115173026	G	T	2
1_955619_G_C	1	955619	rs201073369	G	C	255
1_957568_A_G	1	957568	rs115704555	A	G	2
1_957605_G_A	1	957605	rs756623659	G	A	5

Table continues below

CLNDBN	CLNDSDBID	INTERP
Immunodeficiency_38_with_basal_ganglia_calcification	CN221808:616126	TRUE
Immunodeficiency_38_with_basal_ganglia_calcification	CN221808:616126	TRUE
not_specified	CN169374	FALSE
not_specified	CN169374	FALSE
not_specified	CN169374	FALSE
Congenital_myasthenic_syndrome	C0751882:ORPHA590	TRUE

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/2016-paper-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
```

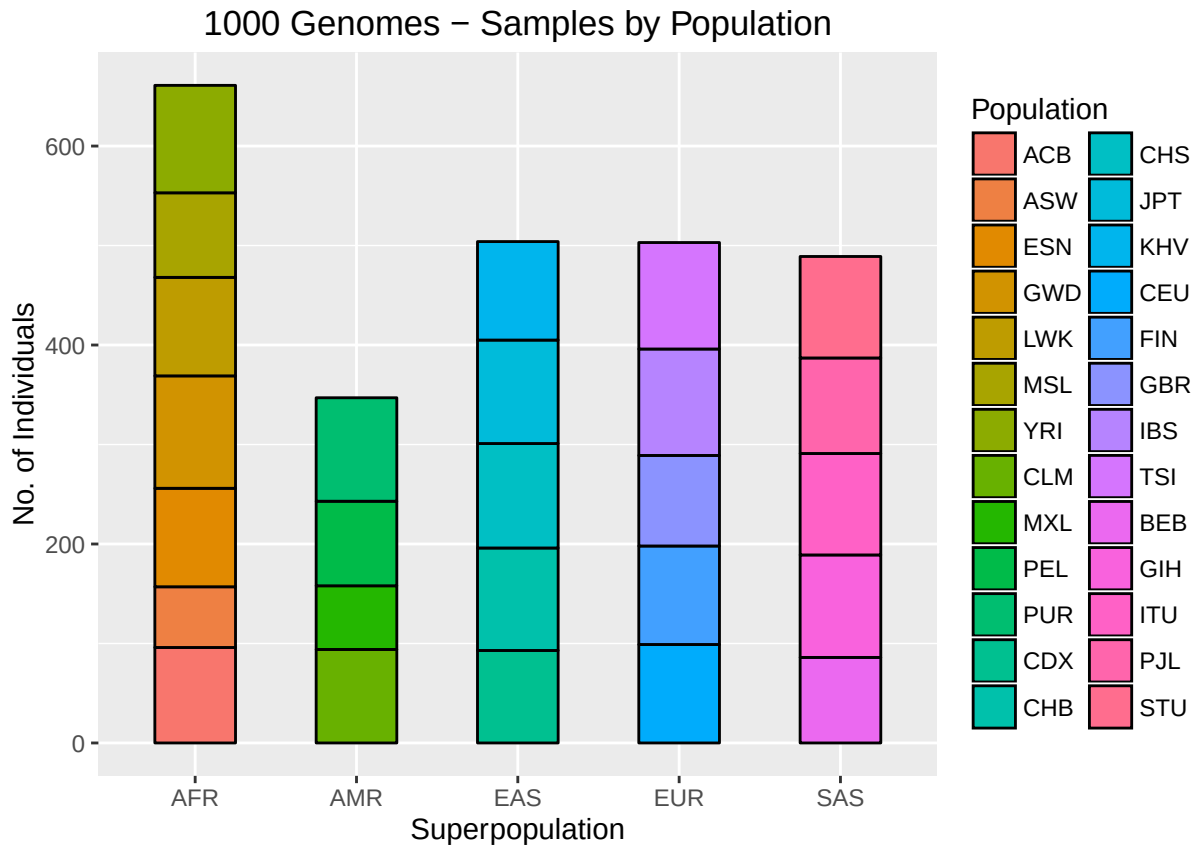
1.4 Collect 1000 Genomes Phase 3 Populations Map

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
HG03082	MSL	AFR	female
HG02666	GWD	AFR	male
HG02419	ACB	AFR	female
HG01396	PUR	AMR	female
NA18967	JPT	EAS	male
HG00650	CHS	EAS	male
NA18532	CHB	EAS	female
HG00269	FIN	EUR	female
NA20832	TSI	EUR	female
HG03716	ITU	SAS	male

```
## Population Distribution
```



1.5 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

1.6 Import and Process 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 ExAC VCFs: 59884 x 45 (selected rows/columns):
```

GENE	AF_EXAC	AF_EXAC_AFR	AF_EXAC_AMR	AF_EXAC_EAS	AF_EXAC_EUR
BRCA1	0.00499800	0.003401361	0.00000000	0	0.0091174325
BRCA1	0.01533000	0.000000000	0.01020408	0	0.0000000000
BRCA1	0.00009098	0.000000000	0.00000000	0	0.0000000000
BRCA1	0.00018100	0.006578947	0.00000000	0	0.0000000000
BRCA1	0.00008700	0.000000000	0.00000000	0	0.0003172589

Table continues below

AF_EXAC_SAS	VAR_ID	CHROM	POS	ID	REF	ALT
0.0036388140	17_41196363_C_T	17	41196363	rs8176320	C	T
0.0222904431	17_41196368_C_T	17	41196368	rs184237074	C	T
0.0001334045	17_41196369_G_T	17	41196369	.	G	T
0.0000000000	17_41196372_T_C	17	41196372	rs189382442	T	C
0.0000000000	17_41196403_A_G	17	41196403	rs182218567	A	G

1.7 Merge ClinVar with 1000 Genomes and ExAC

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

Subset_ClinVar	Number_of_Variants
Total ClinVar	126349
LP/P-ClinVar	33033
LP/P-ClinVar & ACMG	6677
LP/P-ClinVar & ACMG & ExAC	876
LP/P-ClinVar & ACMG & 1000 Genomes	135

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

Subset_1000_Genomes	Number_of_Variants
Total 1000_Genomes & ACMG	141467
1000_Genomes & ACMG & ClinVar	4958
1000_Genomes & ACMG & LP/P-ClinVar	135

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

Subset_ExAC	Number_of_Variants
Total ExAC & ACMG	59884
ExAC & ACMG & ClinVar	10171
ExAC & ACMG & LP/P-ClinVar	876

1.8 Comparison with ClinVar Browser Query Results

`clinvar_query.txt` contains all results matched by the search query: "(APC[GENE] OR MYH11[GENE]... OR WT1[GENE]) AND (clinsig_pathogenic[prop] OR clinsig_likely_pathogenic[prop])" from the ClinVar website. The exact query is saved in /Supplementary_Files/query_input.txt
This presents another way of collecting data from ClinVar.

Intermediate step: convert hg38 locations to hg19 using the Batch Coordinate Conversion tool (liftOver) from UCSC Genome Browser Utilities.

```
## ClinVar Query Results Table (substitutions only): 6714 x 13 (selected rows/columns)
```

VAR_ID	Gene(s)	Condition(s)	Frequency
X_100652891_C_G	GLA	Fabry disease	GMAF:0.00050(G)
11_47374186_C_G	MYBPC3	Primary familial hypertrophic cardiomyopathy	GMAF:0.00020(G)
11_47355233_C_G	MYBPC3	Familial hypertrophic cardiomyopathy 4	GMAF:0.00020(G)
11_47364162_C_G	MYBPC3	Familial hypertrophic cardiomyopathy 4	GMAF:0.00020(G)
14_23886482_G_C	MYH7	not specified	GMAF:0.00020(C)
14_23893148_C_G	MYH7	Primary dilated cardiomyopathy	GO-ESP:0.00046(G)
1_17355075_A_T	SDHB	Gastrointestinal stromal tumor	GMAF:0.00120(T)
1_17380507_G_C	SDHB	Cowden syndrome 2	GO-ESP:0.01323(C)

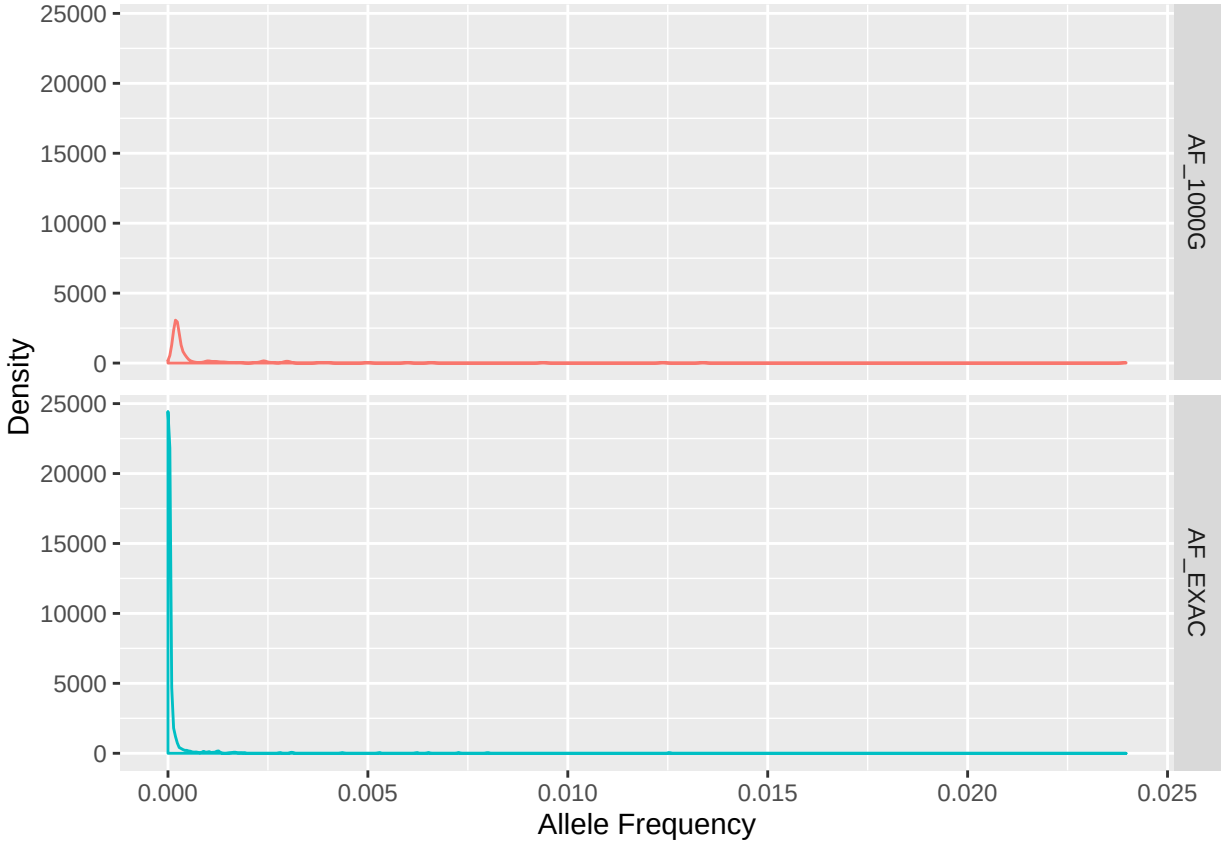
```
## Breakdown of ClinVar Query Results Table:
```

Subset	Number_of_Variants
Initial Count	12525
Filter Substitutions (N>N')	6732
Filter Coupling/Bad-Locations	6714
In ClinVar VCF	509
In LP/P-ClinVar VCF	503
^ & ACMG & ExAC	49
^ & ACMG & 1000 Genomes	9
^ & ACMG & ExAC & 1000 Genomes	8

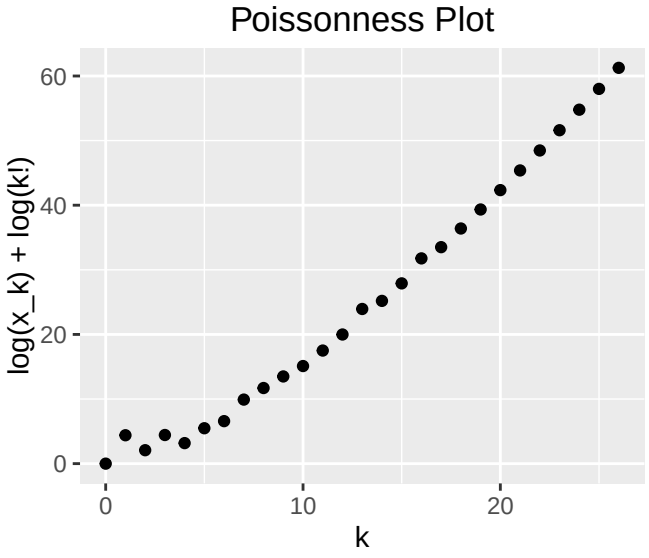
```
## Note the 12-fold reduction after merging the online query results with the VCF.
```

2 Plot Summary Statistics Across Populations

2.1 Distribution of Allele Frequencies



The distribution of allele frequencies is approximately Poisson, with "Poissonness plot" correlation = 0.99. The Poissonness plot (Hoaglin 1980) is defined as the plot of $log(x_k) + log(k!)$ vs. k, as shown below:



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.

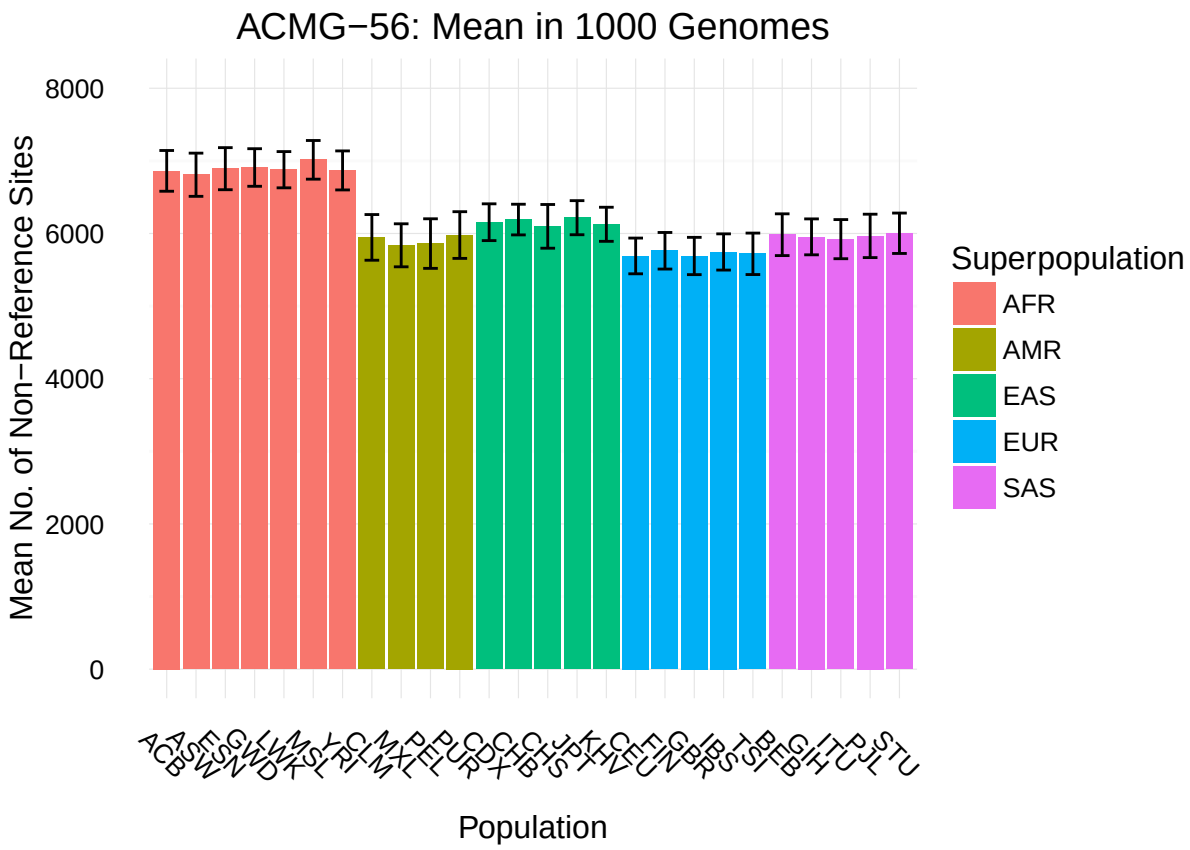
Ex: the genotype of 3 variants in 3 people looks like this:

	HG00097	HG00099	HG00100
Variant 1	0	0	0
Variant 2	0	0	0
Variant 3	0	0	0

Count the number of non-reference sites per individual:

HG00097	HG00099	HG00100
0	0	0

```
## Mean = 0
```



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

2.2.0.2 For 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:

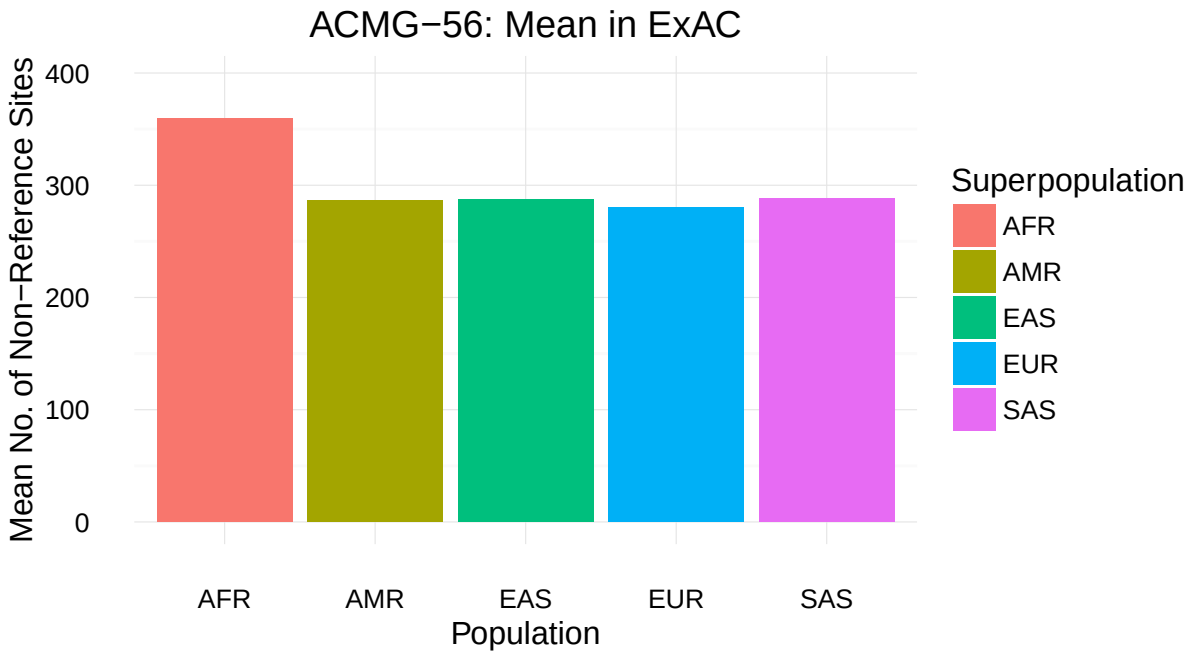
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

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 $\sum E(V_i) = \sum(columns)$.

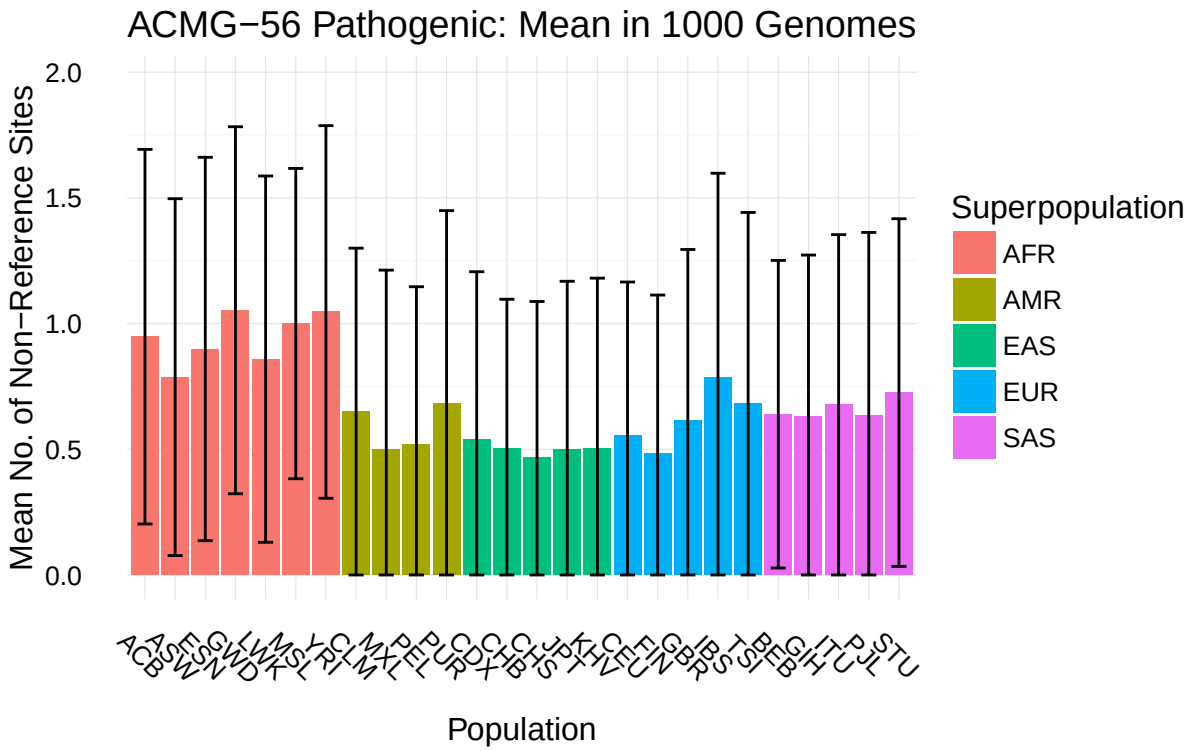
AFR	AMR	EAS	EUR	SAS
0.55	0.36	0.51	0	0.7

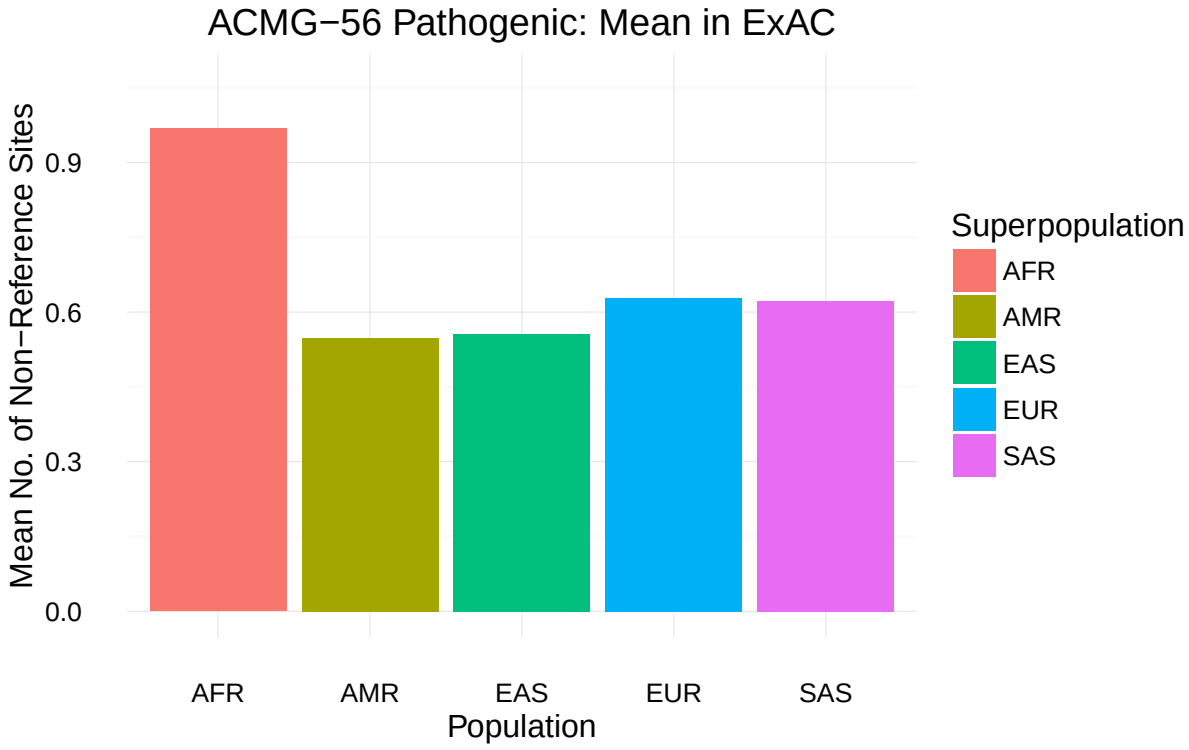


2.3 Pathogenic Non-Reference Sites

2.3.0.1 For 1000 Genomes and ExAC

This is the same procedure as above, but performed only on the subset of variants that are pathogenic.





2.4 Fraction of Individuals with Pathogenic Sites

2.4.0.1 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-56 genes.

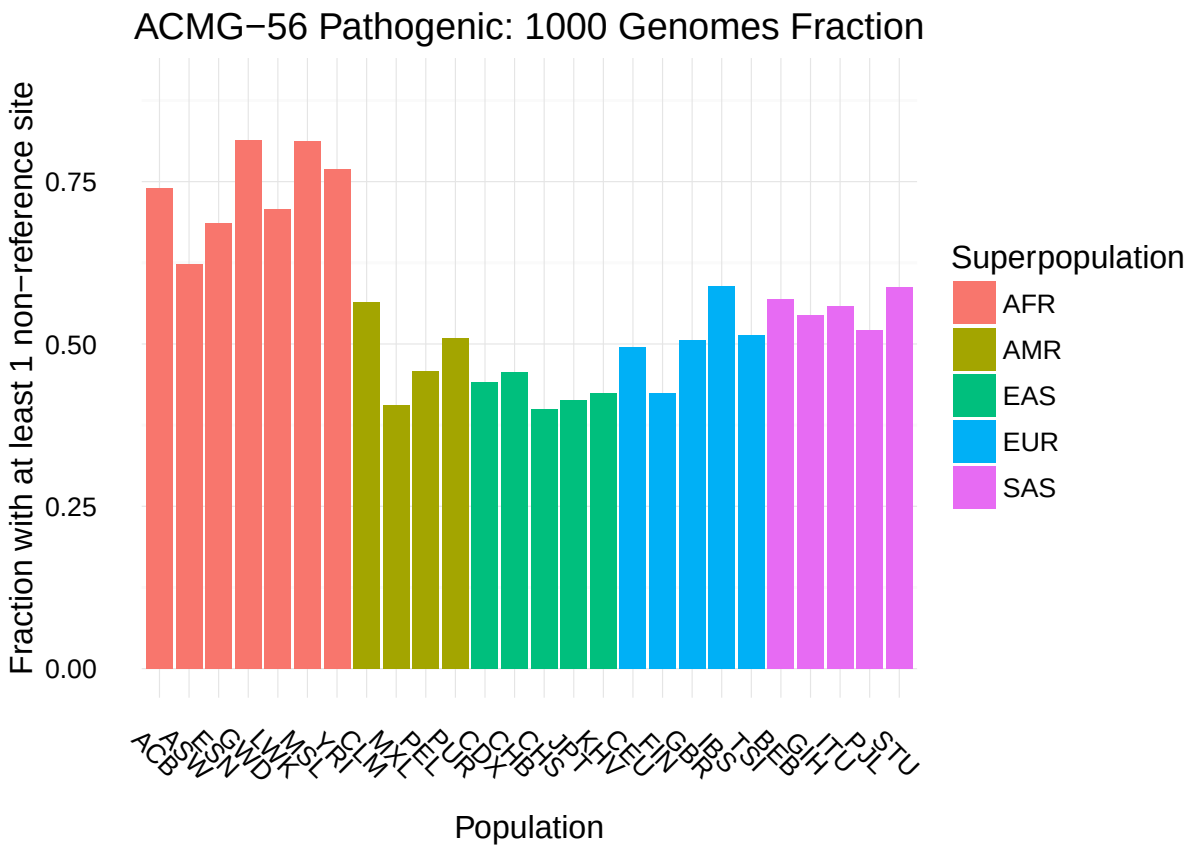
Ex: the genotype of 3 variants in 3 people looks like this:

	HG00097	HG00099	HG00100
Variant 1	0	0	0
Variant 2	0	0	0
Variant 3	0	0	0

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

HG00097	HG00099	HG00100
0	0	0

```
## Mean = 0
```



2.4.0.2 For 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:

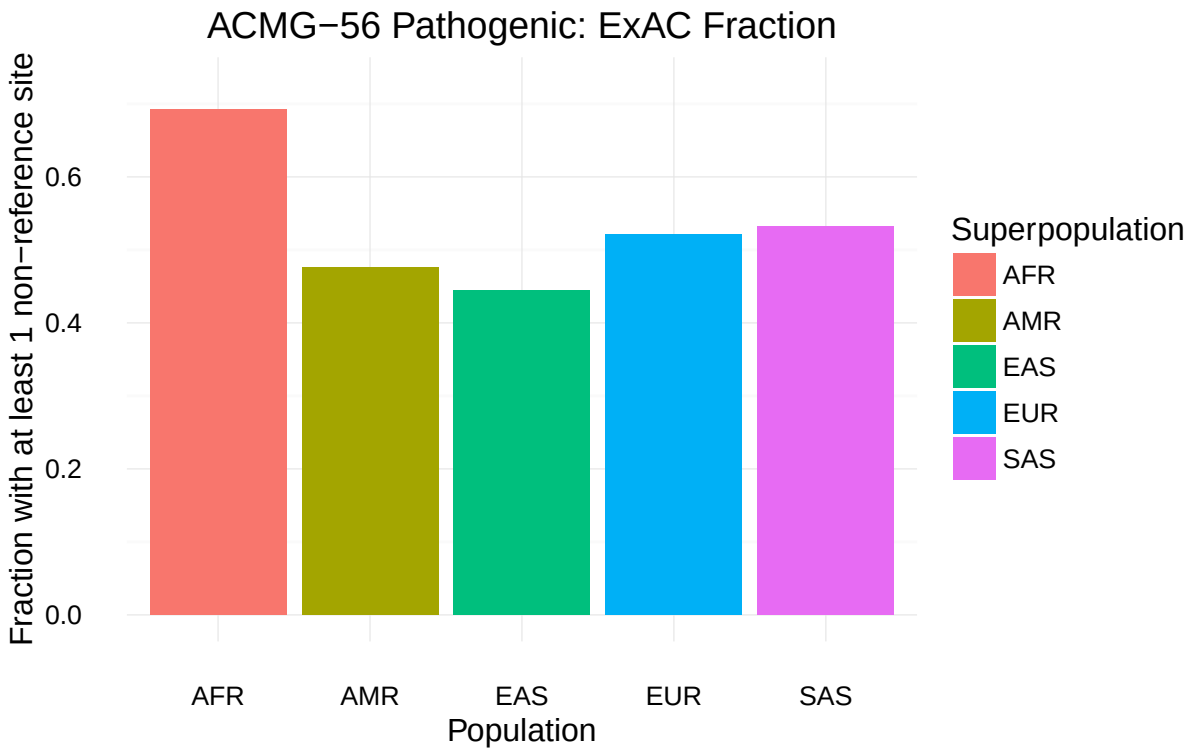
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

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

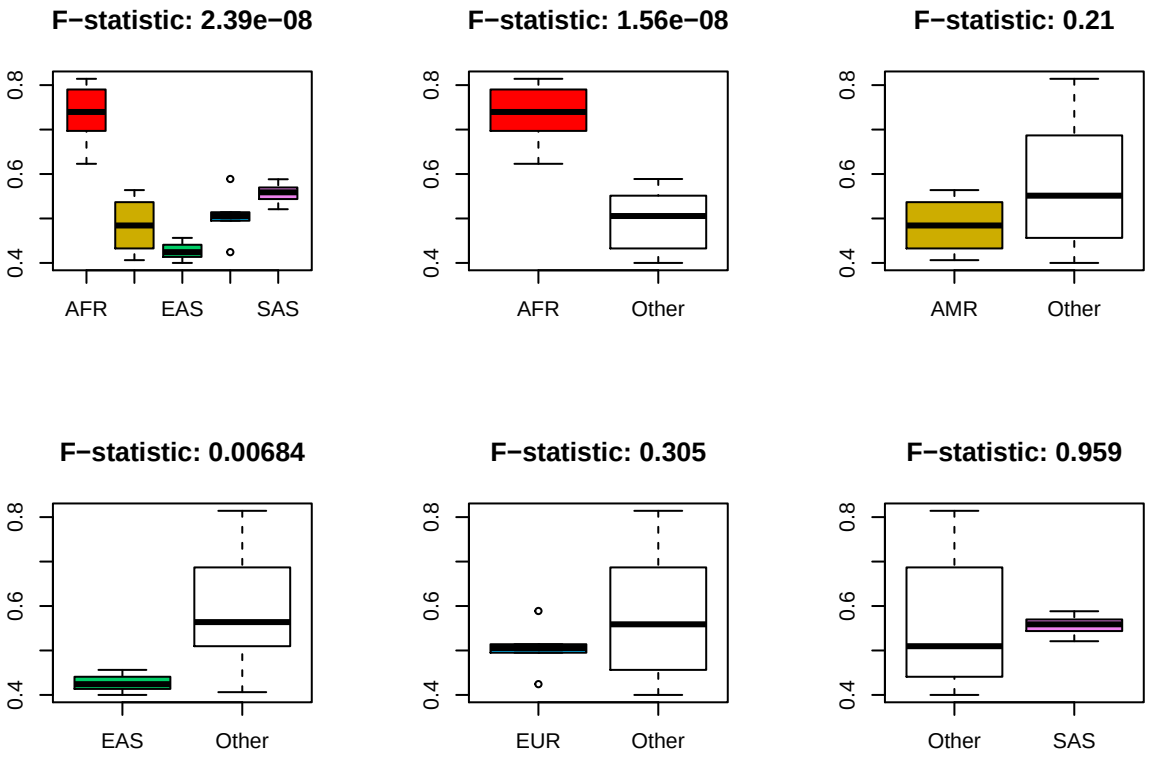
The expected (mean) number of non-reference sites is given by $1 - \prod(1 - AF)^2$.

AFR	AMR	EAS	EUR	SAS
0.4816	0.36	0.51	0	0.6031

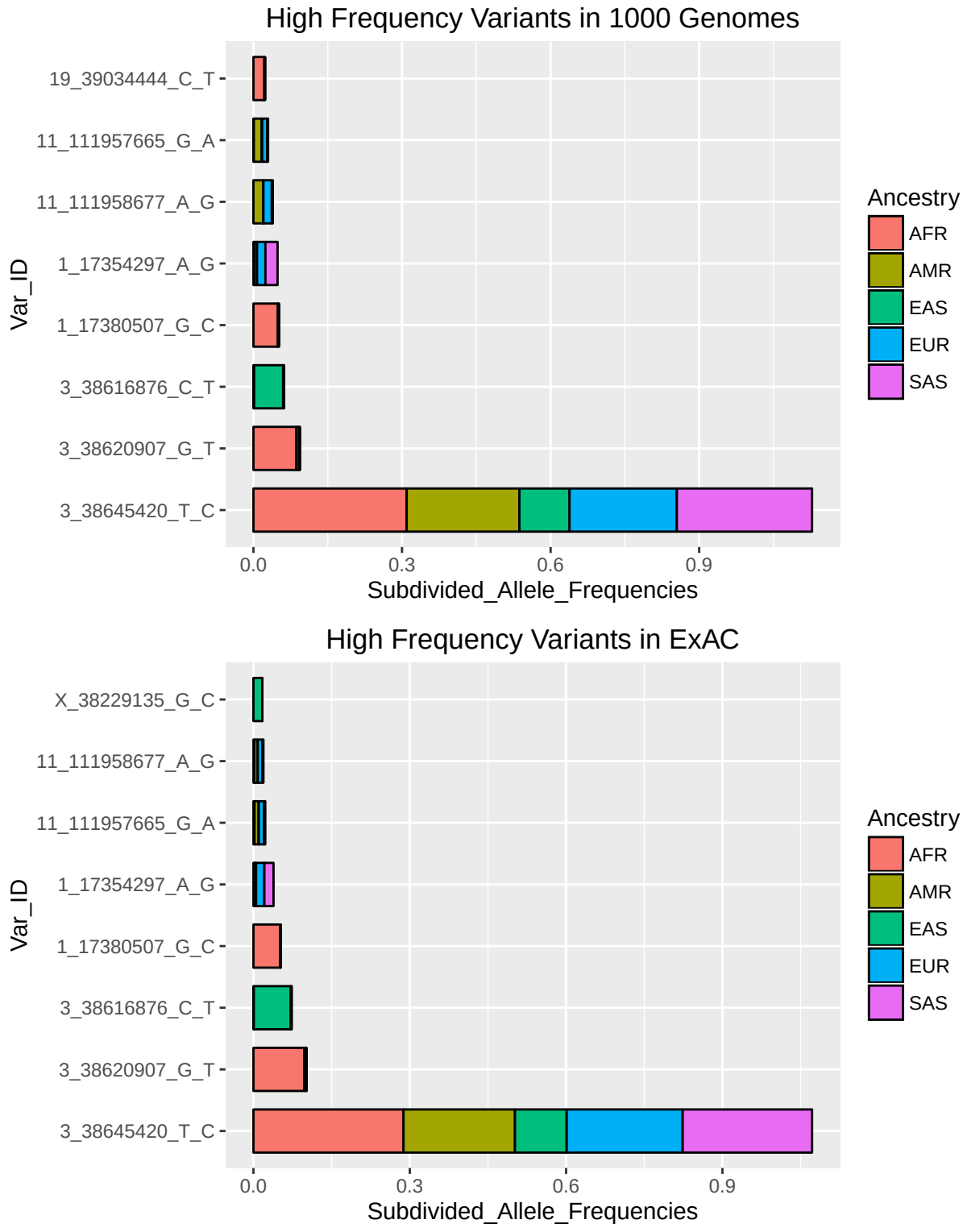


2.5 Test Statistics for Ancestral Differences

F-statistic/T-statistic: probability that the different groups are sampled from distributions with the same mean. These plots are from 4(a) - 1000 Genomes Fraction with 1+ Non-Reference Site, but can be replicated for plots 2(ab) and 3(ab) as well.



2.6 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. Allele Frequency = $P(V_x)$
3. Allelic Heterogeneity = $P(V_x|D_x)$
4. Penetrance = $P(D_x|V_x)$

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

$$P(D_x|V_x) = \frac{P(D_x) * P(V_x|D_x)}{P(V_x)} = \frac{Prevalence * Allelic.Heteogeneity}{Allele.Frequency}$$

To compute penetrance estimates for each of the diseases related to the ACMG-56 genes, we will use the prevalence data we collected into `Literature_Prevalence_Estimates.csv`, allele frequency data from 1000 Genomes and ExAC, and a broad range of values for allelic heterogeneity.

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 56 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.

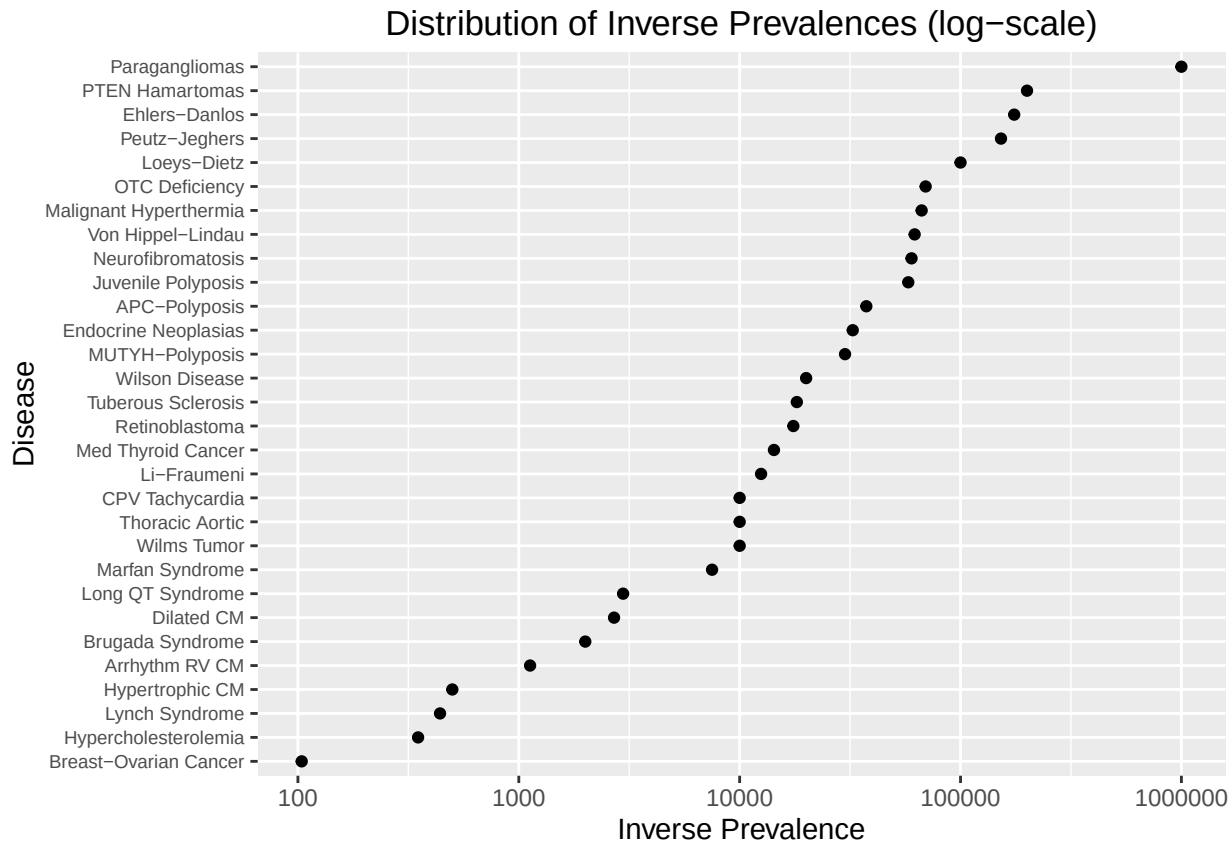
`## Table of Literature-Based Estimates of Disease Prevalence 30 x 15 (selected rows/columns):`

Gene	Phenotype	Abbreviation
MLH1|MSH2|MSH6|PMS2	Lynch syndrome	HNPCC
APC	Familial adenomatous polyposis	HCRC-AD
VHL	Von Hippel-Lindau syndrome	VHL
TSC1|TSC2	Tuberous sclerosis complex	TSC

Table continues below

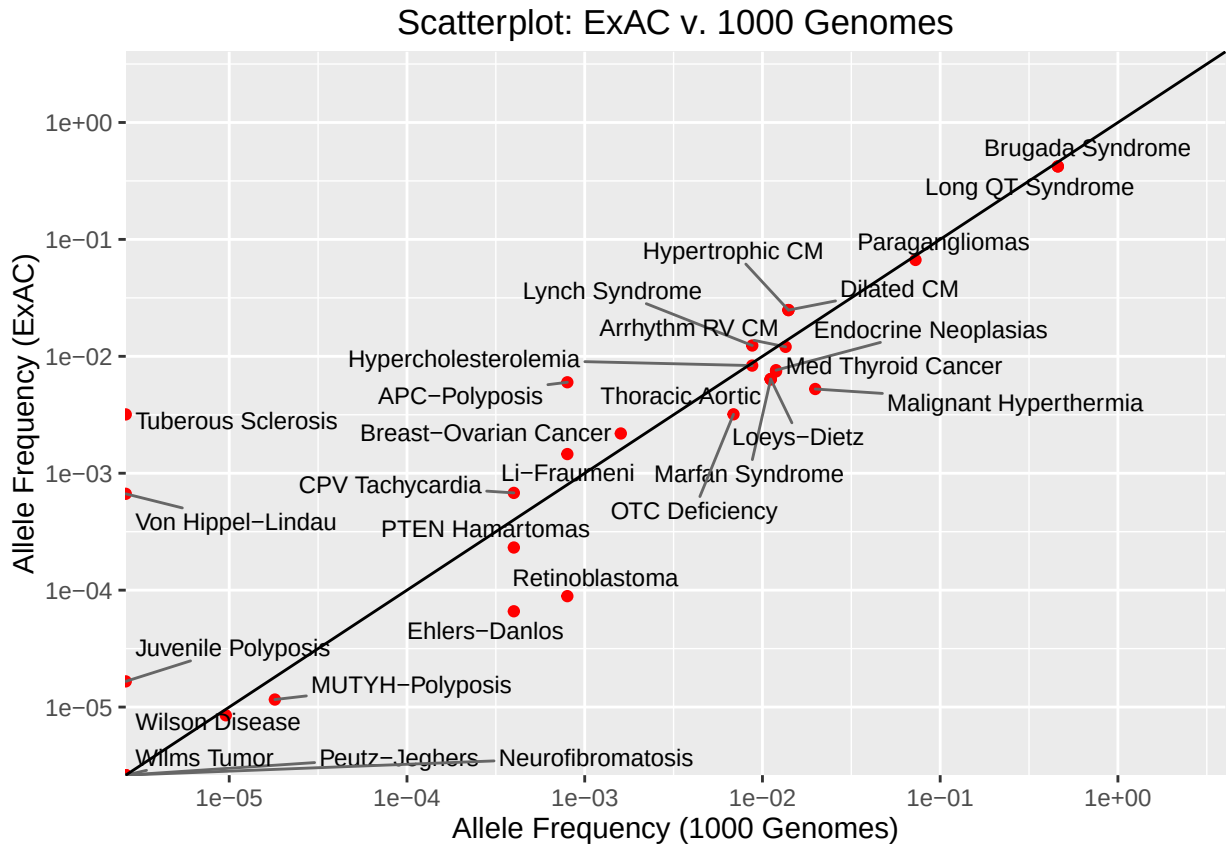
Inverse_Prevalence
440
31250-43668
38951-85000
11300-25000

3.3 Distribution of Prevalences

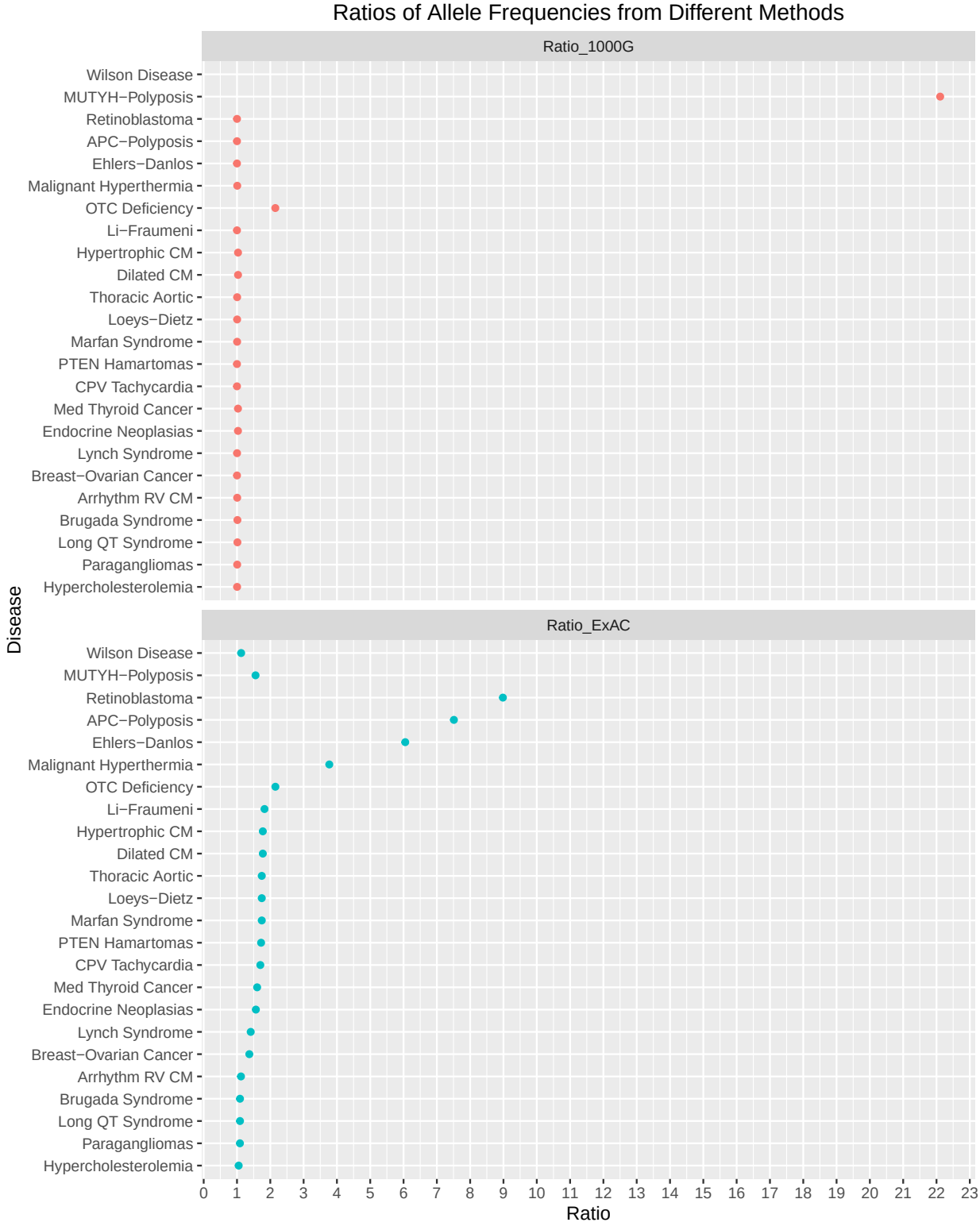


3.4 Collect and Aggregate Allele Frequencies at the Disease-Level

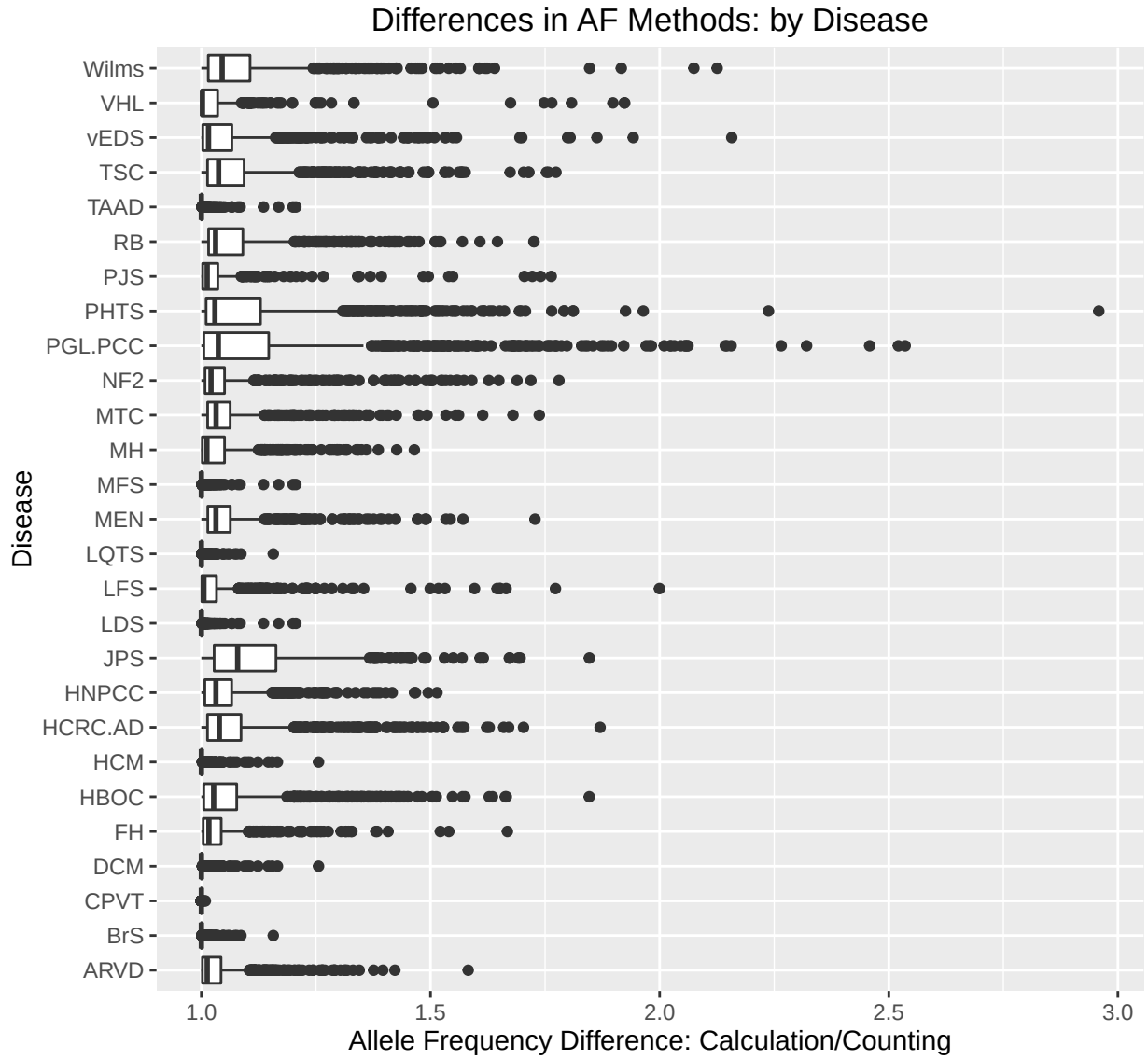
We define AF(disease) as the probability of having at least 1 variant associated with the disease.
The frequencies across the relevant variants can be aggregated in two ways:
(1) By direct counting, from genotype data in 1000 Genomes.
(2) AF(disease) $= 1 - \prod_{variant}(1 - AF_{variant})$, from population data in ExAC (assumes independence).

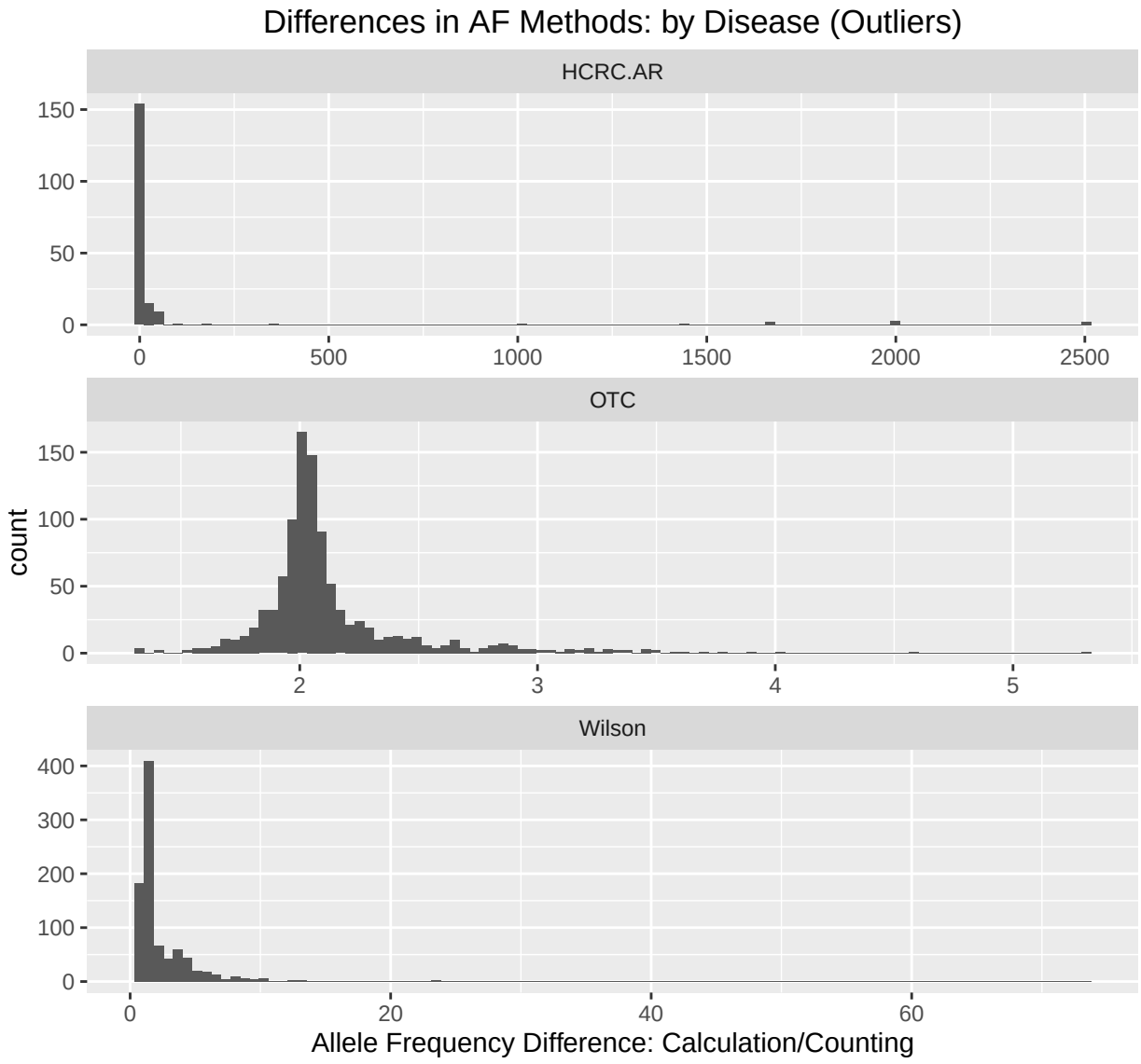


Ratio_1000G (red, top) computes AF(calculation in 1000 Genomes) / AF(counting in 1000 Genomes).
Ratio_ExAC (blue, bottom) computes AF(calculation in ExAC) / AF(calculation in 1000 Genomes).



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).



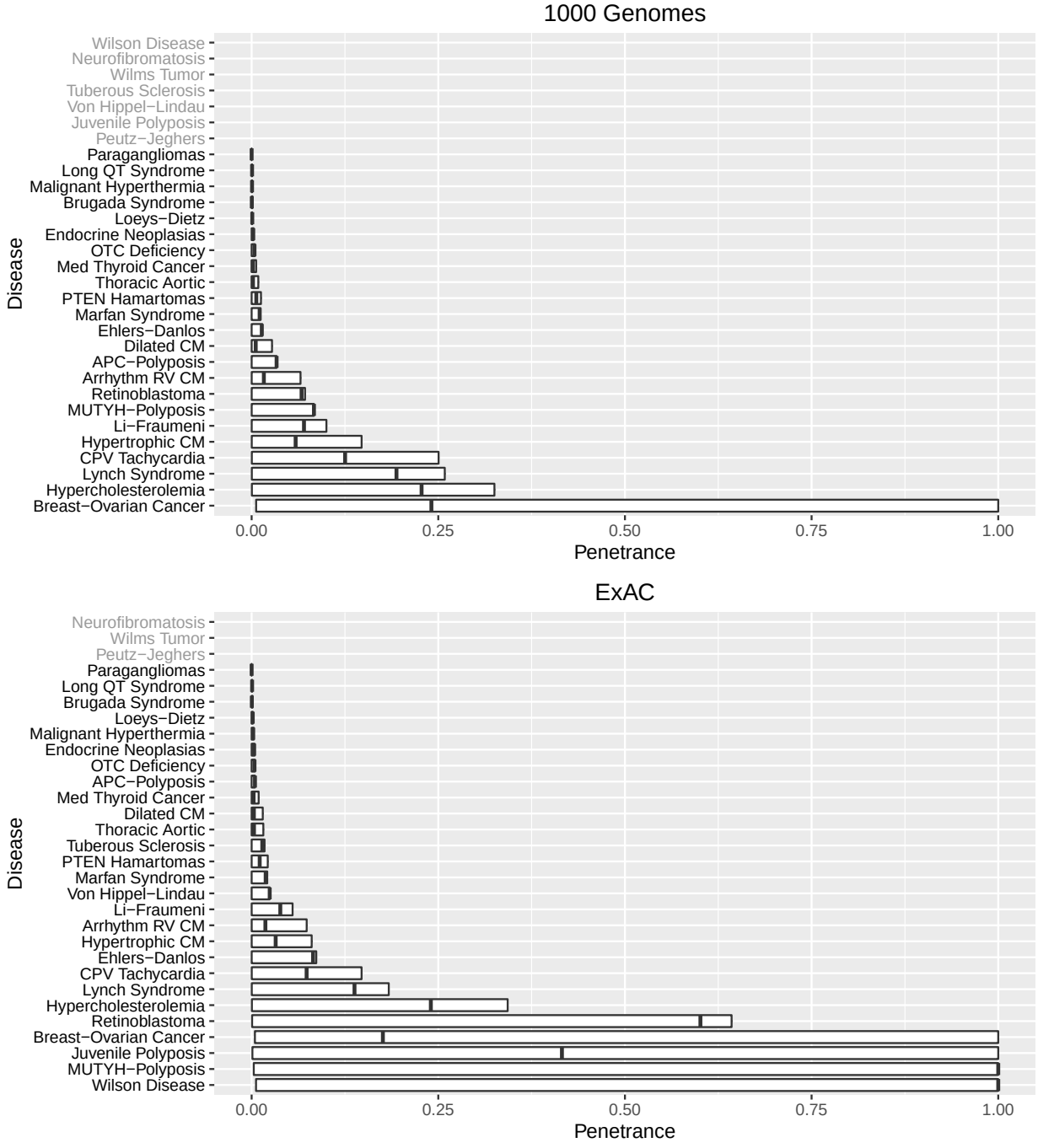

Differences in AF Methods: by Disease (Outliers)
HCRC.AR
OTC
Wilson
count
Allele Frequency Difference: Calculation/Counting


```
## 30 diseases x 1000 points = 30,000 points.This plot has been downsampled 10x and contains 3,000 poin
```

Pearson correlation: 0.99
Mean ratio (Calculation/Counting): 1.07

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) = 0.5,
the right end of the boxplot indicates P(V|D) = 1.

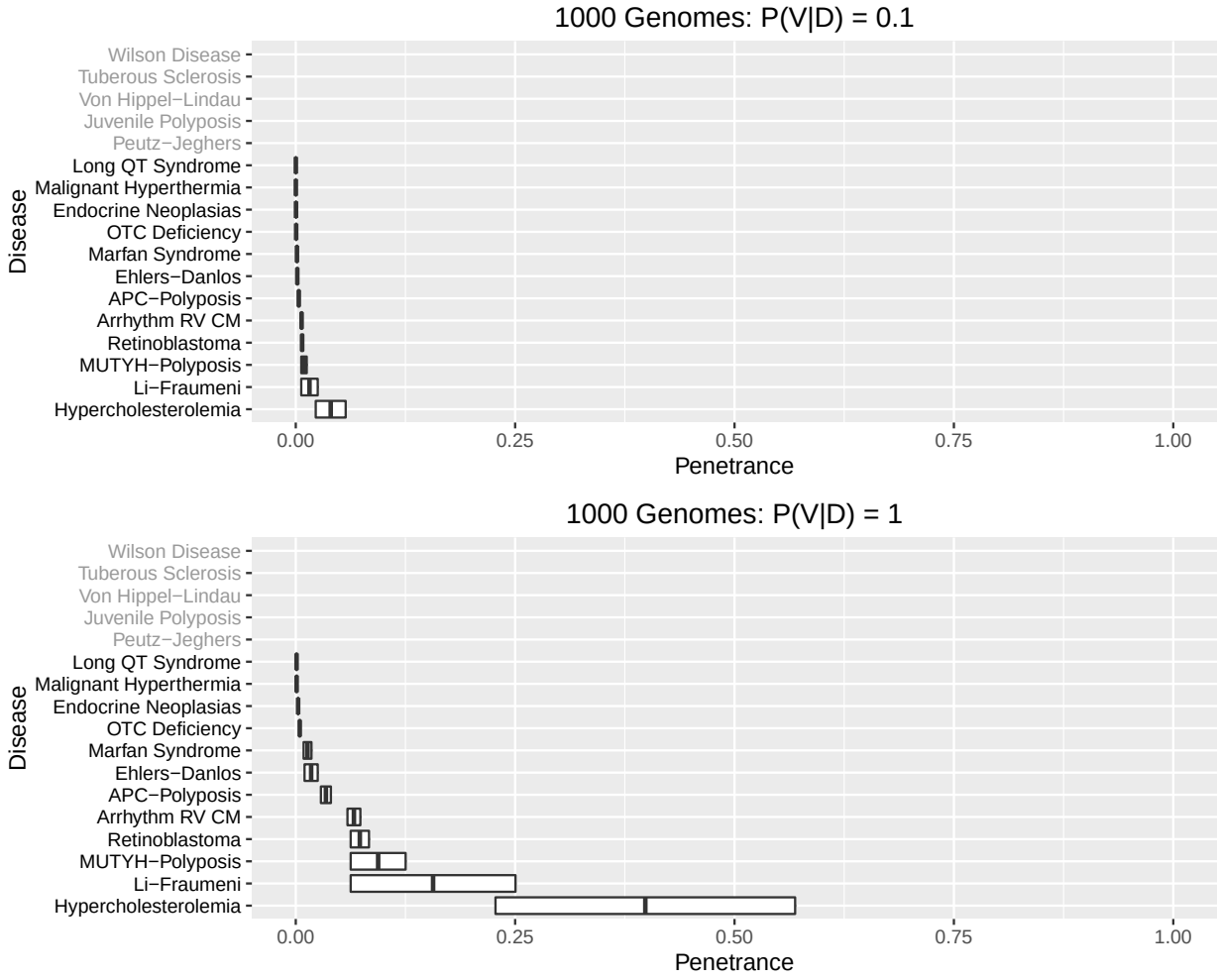


Note 1: the grayed-out empty lines at the top all indicate no allele frequency (disease_AF) data.
Note 2: For breast-ovarian cancer, mean theoretical penetrance > 1. This is because the assumed allelic heterogeneity (0.25) is greater than is possible, given the empirical prevalence and allele frequencies.

3.6 Penetrance as a Function of P(D)

Disease	Prevalence_Ratio
Peutz-Jeghers	0.1
Li-Fraumeni	0.2
Juvenile Polyposis	0.2
Wilson Disease	0.3
Malignant Hyperthermia	0.3
Ehlers-Danlos	0.4
Long QT Syndrome	0.4
Hypercholesterolemia	0.4
MUTYH-Polyposis	0.5
Von Hippel-Lindau	0.5
Tuberous Sclerosis	0.5
Marfan Syndrome	0.5
APC-Polyposis	0.7
Retinoblastoma	0.8
Arrhythm RV CM	0.8
OTC Deficiency	0.8
Endocrine Neoplasias	0.9

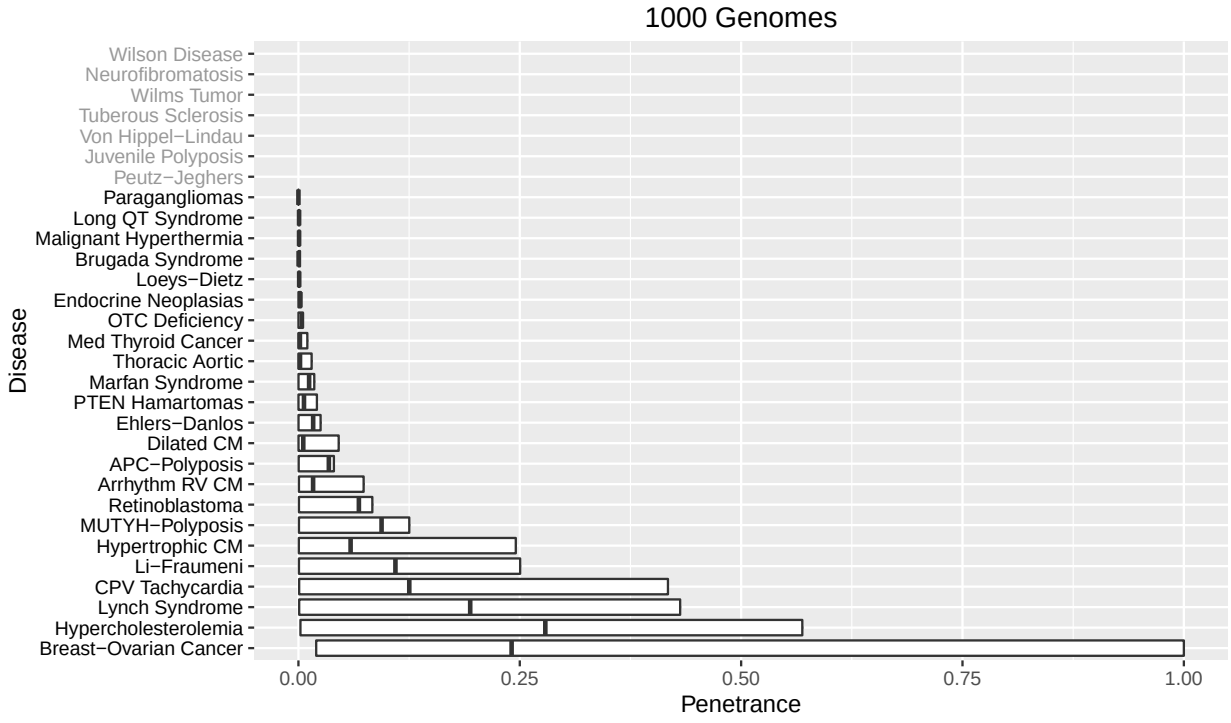
The left end of the boxplot indicates P(D) = upper value,
the bold line in the middle indicates P(D) = mean(values),
the right end of the boxplot indicates P(D) = lower value.

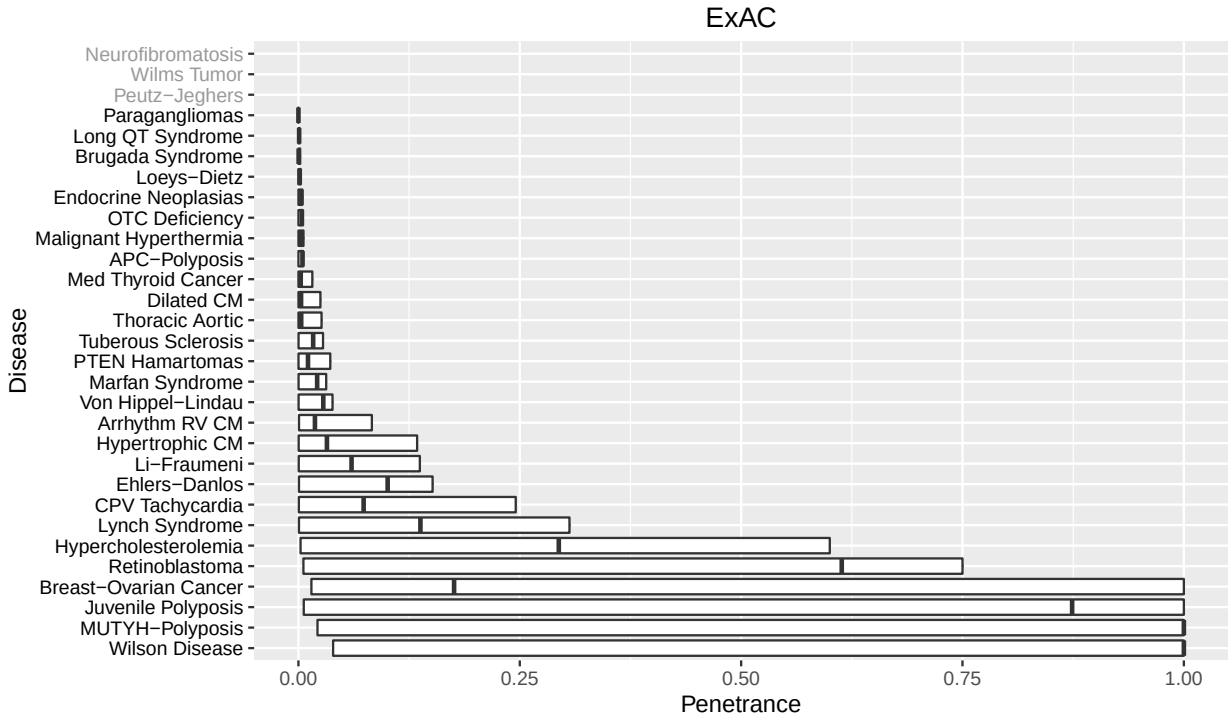


This can only be computed in the 9 cases where a prevalence range was given (rather than a point estimate) and the disease-level allele frequency is known (in this plot: all of them except Puetz-Jeghers).

3.7 Max/Min Penetrance as a Function of P(D) and P(V|D)

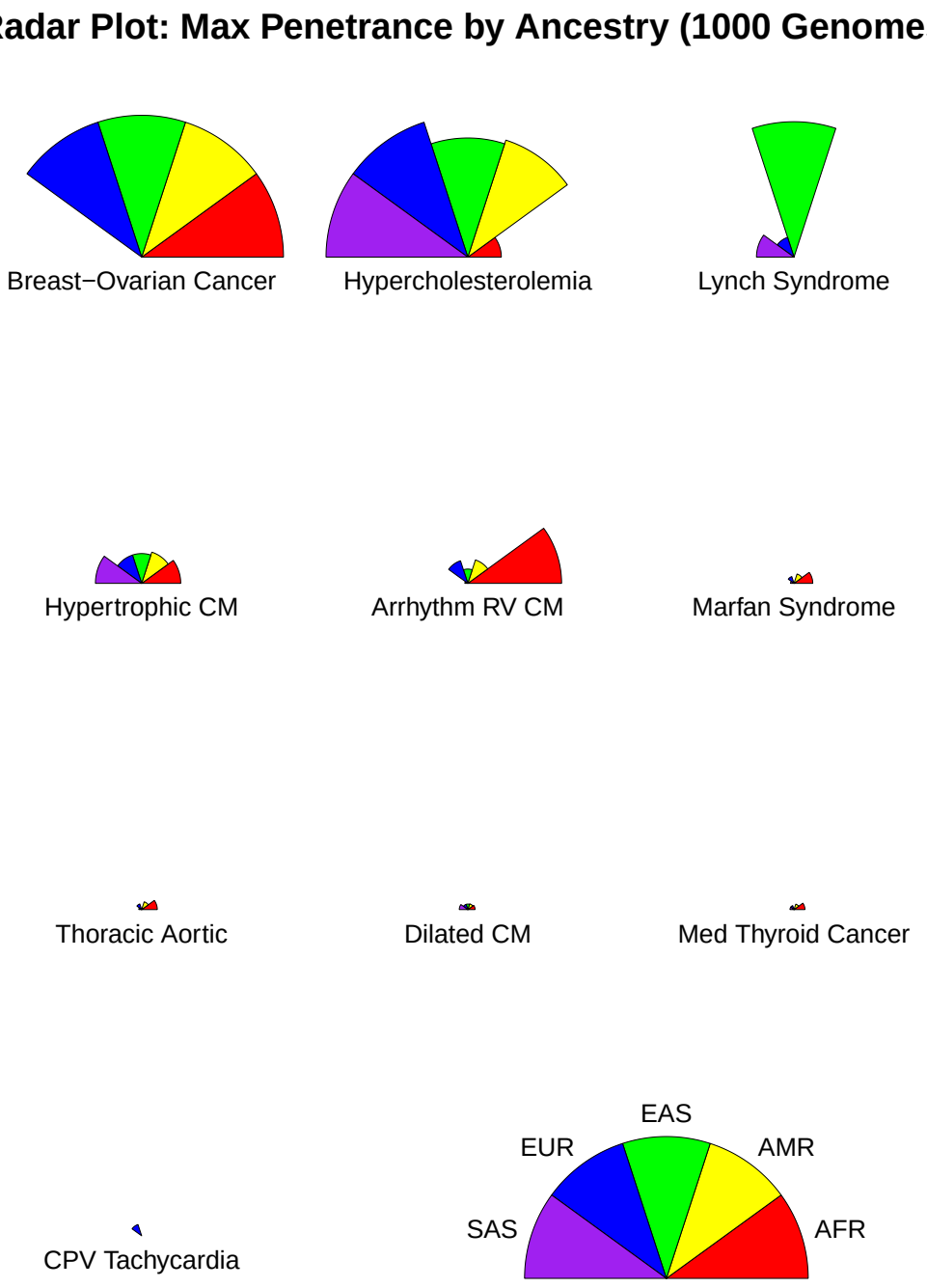
The left end of the boxplot indicates P(D) AND P(V|D) = lower value,
the bold line in the middle indicates P(D) AND P(V|D) = mean(values),
the right end of the boxplot indicates P(D) AND P(V|D) = upper value.



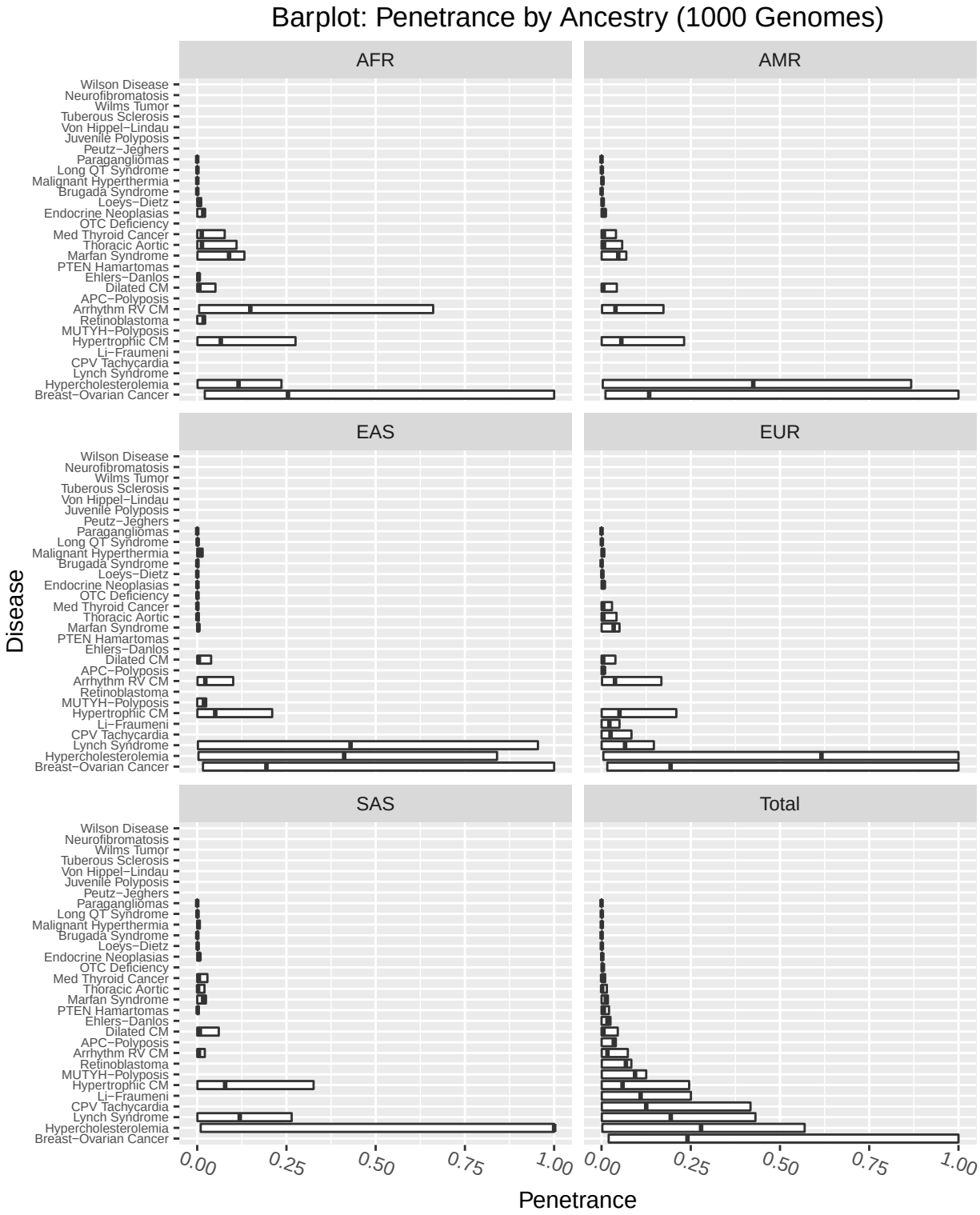


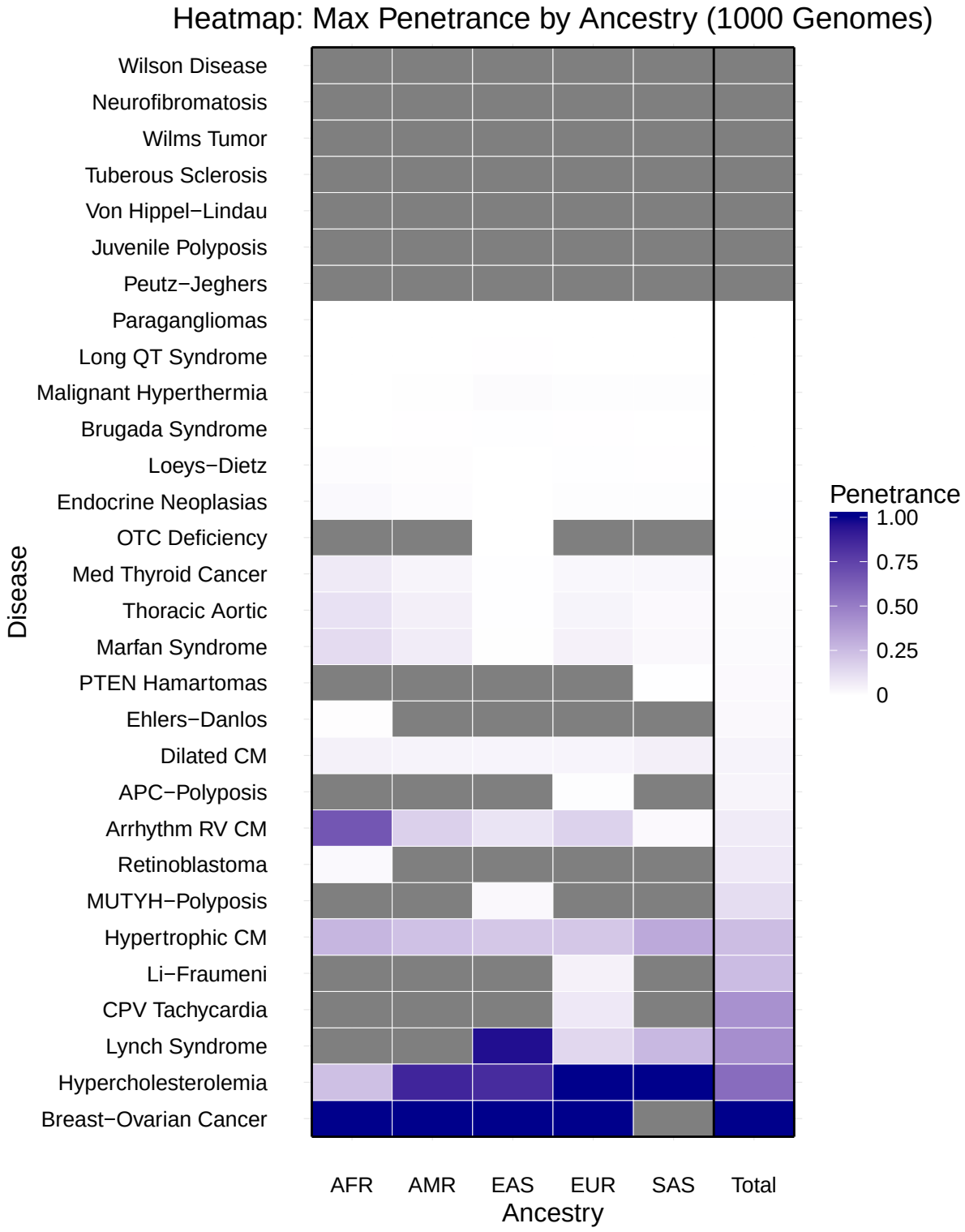
Note: Prevalence ranges of 5x were assumed for all point estimates of prevalence.
For example: a point estimate of 0.3 would be given the range [0.1, 0.5].

3.8 Penetrance Estimates by Ancestry



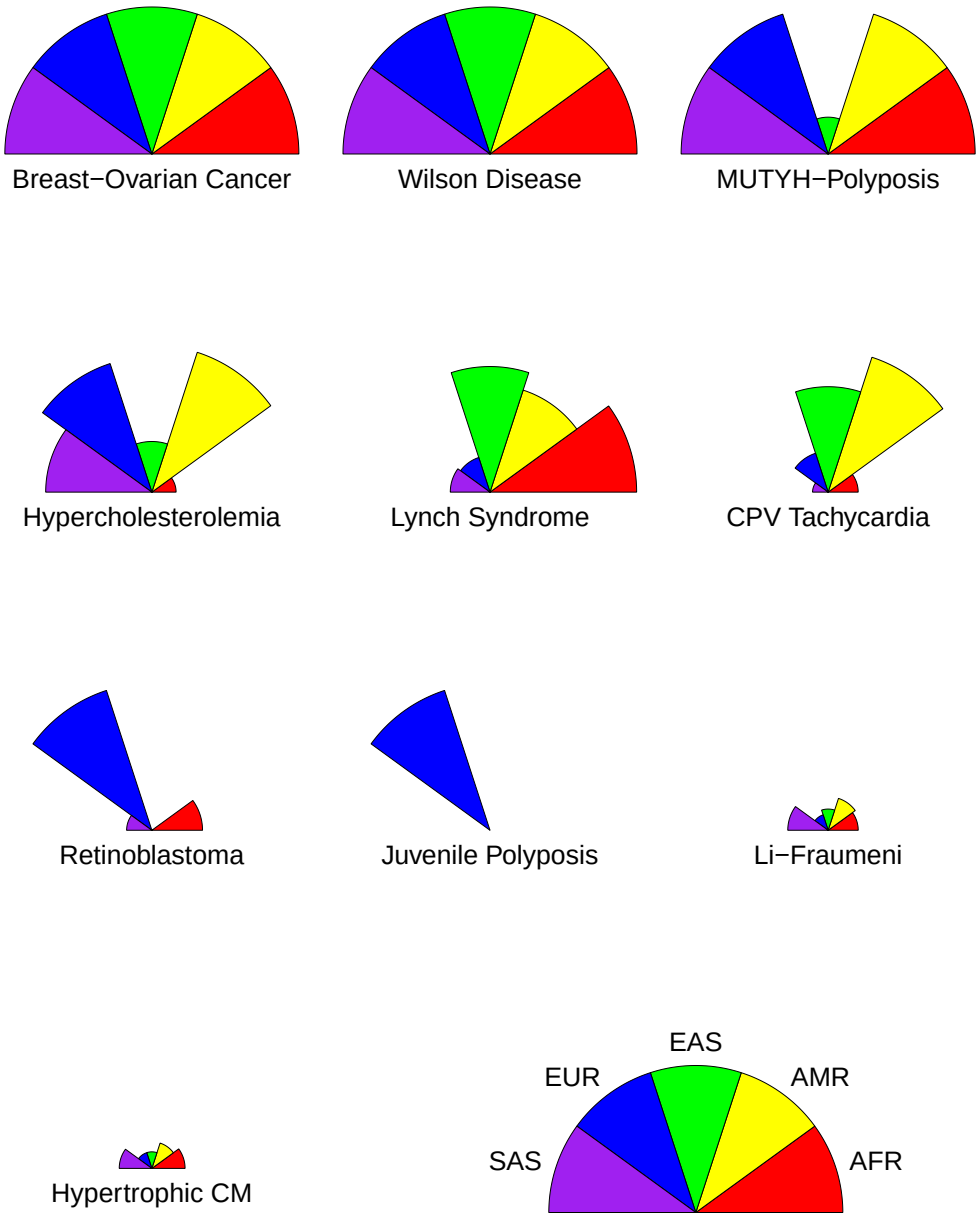

```
## [1] These are the top 10 diseases by summed allele frequencies. NULL values are not plotted.
## [1] Each radius is proportional to the penetrance of the disease in the given population.
```


Barplot: Penetrance by Ancestry (1000 Genomes)
AFR
AMR
EAS
EUR
SAS
Total
Disease
Wilson Disease
Neurofibromatosis
Wilms Tumor
Tuberous Sclerosis
Von Hippel−Lindau
Juvenile Polyposis
Peutz−Jeghers
Paragangliomas
Long QT Syndrome
Malignant Hyperthermia
Brugada Syndrome
Loeys−Dietz
Endocrine Neoplasias
OTC Deficiency
Med Thyroid Cancer
Thoracic Aortic
Marfan Syndrome
PTEN Hamartomas
Ehlers−Danlos
Dilated CM
APC−Polyposis
Arrhythm RV CM
Retinoblastoma
MUTYH−Polyposis
Hypertrophic CM
Li−Fraumeni
CPV Tachycardia
Lynch Syndrome
Hypercholesterolemia
Breast−Ovarian Cancer
0.00
0.25
0.50
0.75
1.00
Penetrance

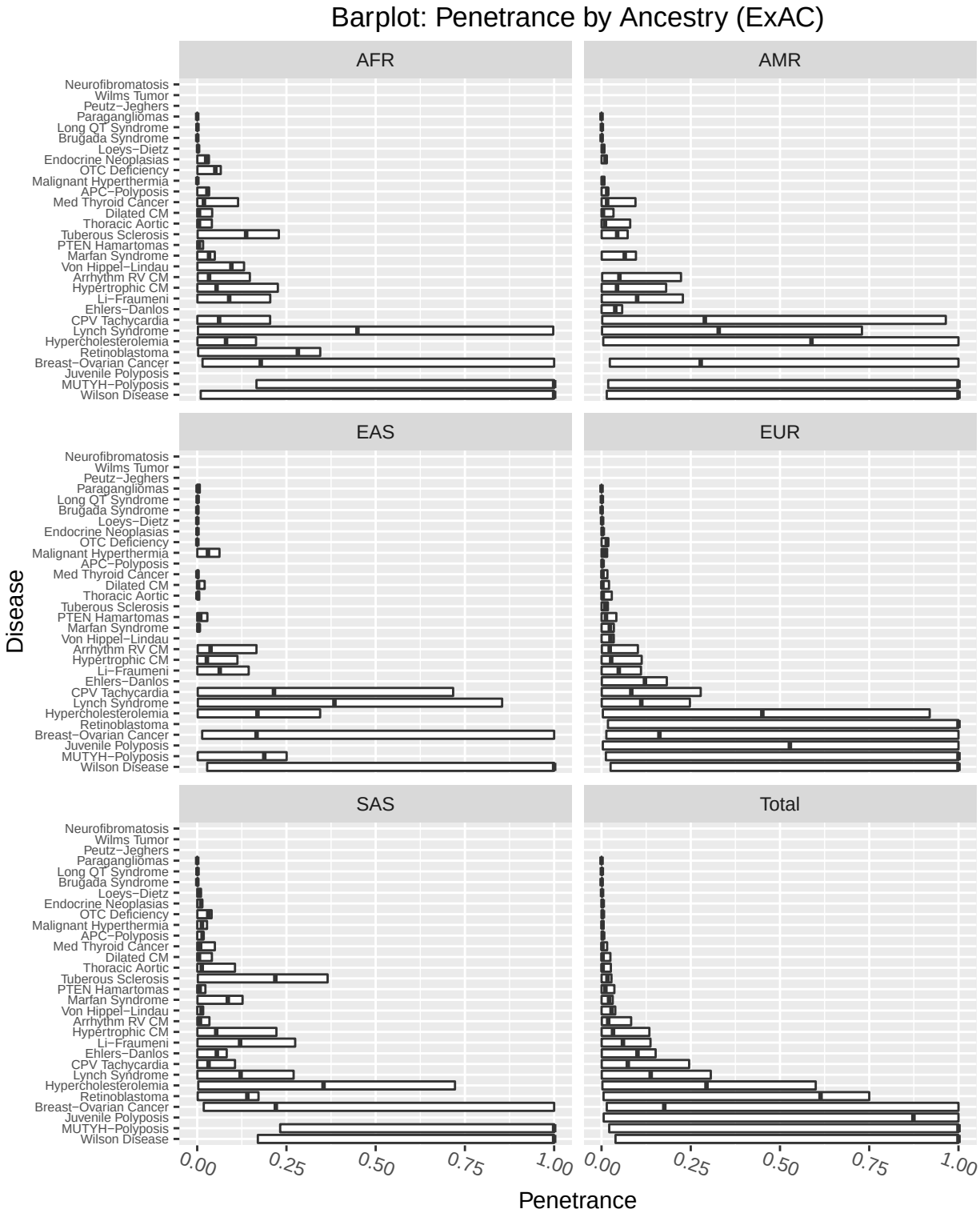


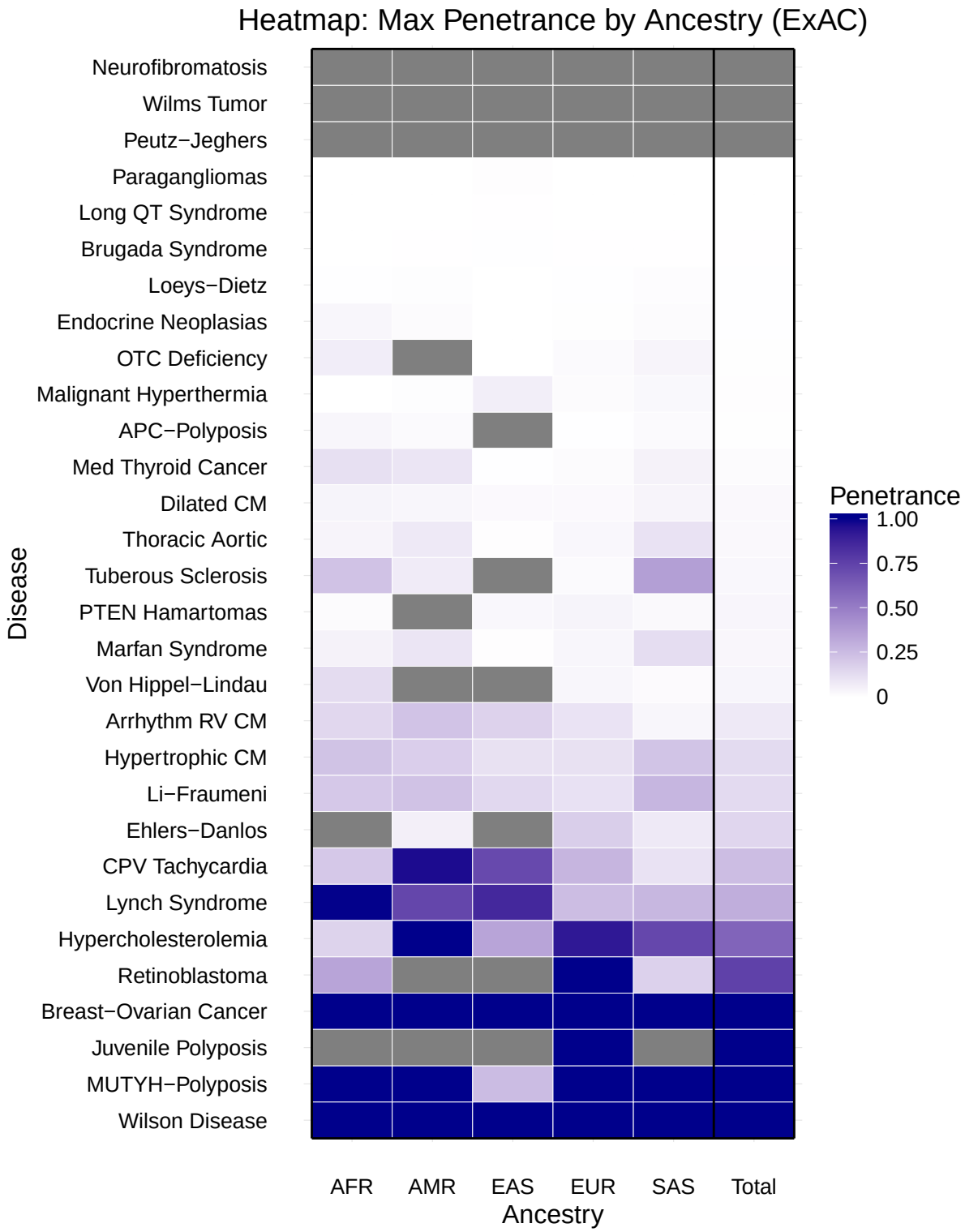
Dark gray boxes are NA: no associated variants discovered in that ancestral population.

Radar Plot: Max Penetrance by Ancestry (ExAC)



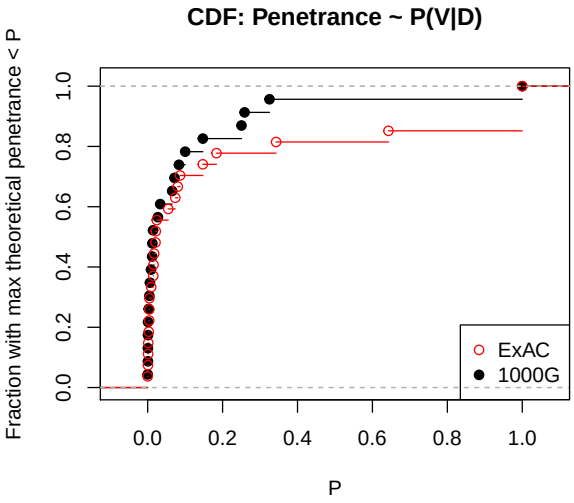

```
## [1] These are the top 10 diseases by summed allele frequencies. NULL values are not plotted.
## [1] Each radius is proportional to the penetrance of the disease in the given population.
```

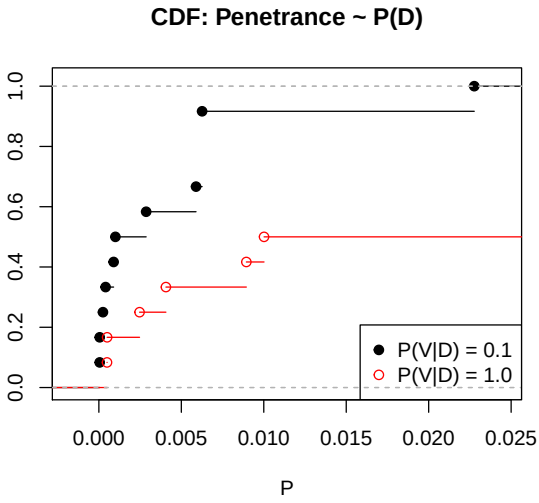

Barplot: Penetrance by Ancestry (ExAC)
AFR
AMR
EAS
EUR
SAS
Total
Disease
Neurofibromatosis
Wilms Tumor
Peutz−Jeghers
Paragangliomas
Long QT Syndrome
Brugada Syndrome
Loeys−Dietz
Endocrine Neoplasias
OTC Deficiency
Malignant Hyperthermia
APC−Polyposis
Med Thyroid Cancer
Dilated CM
Thoracic Aortic
Tuberous Sclerosis
PTEN Hamartomas
Marfan Syndrome
Von Hippel−Lindau
Arrhythm RV CM
Hypertrophic CM
Li−Fraumeni
Ehlers−Danlos
CPV Tachycardia
Lynch Syndrome
Hypercholesterolemia
Retinoblastoma
Breast−Ovarian Cancer
Juvenile Polyposis
MUTYH−Polyposis
Wilson Disease
0.00
0.25
0.50
0.75
1.00
Penetrance

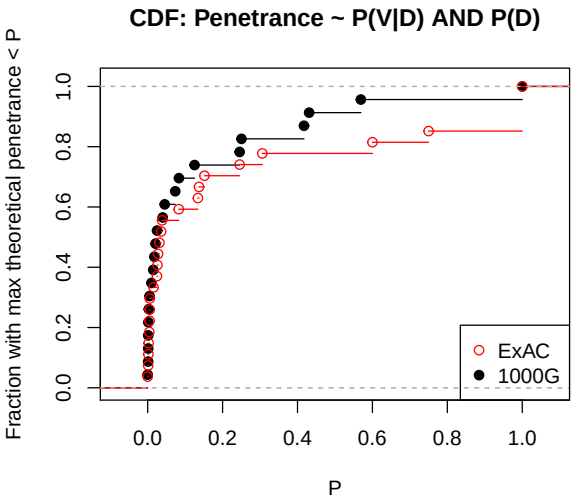


Dark gray boxes are NA: no associated variants discovered in that ancestral population.

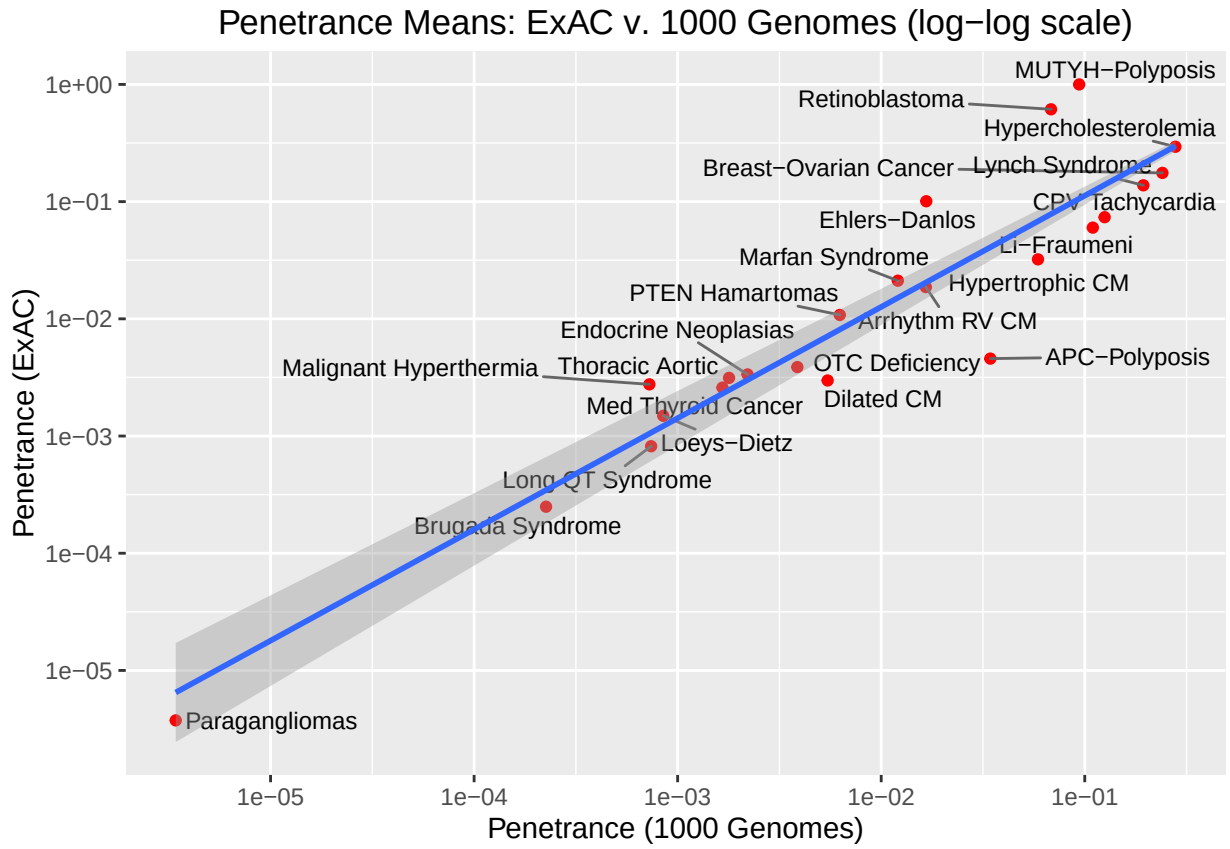
3.9 Empirical CDFs for All Penetrance Plots







3.10 Comparing Mean Penetrance between ExAC and 1000 Genomes



The Pearson correlation is 0.38.
Max penetrance values computed using 1000 Genomes are 0.132-fold larger than those computed using ExAC.