Cardiac ACMG-ClinVar Penetrance Estimation

James Diao, under the supervision of Arjun Manrai

June 27, 2017

Contents

1 Download, Transform, and Load Data **2**

1.1 Collect ACMG Gene Panel . . . 2

1.2 Download ClinVar VCF . . . 3

1.3 Download 1000 Genomes VCFs . . . 3

1.4 Import and Process 1000 Genomes VCFs . . . 4

1.5 Import and Process gnomAD/ExAC VCFs . . . 4

1.6 Collect 1000 Genomes Phase 3 Populations Map . . . 5

1.7 Merge ClinVar with gnomAD, ExAC, and 1000 Genomes . . . 5

1.8 Overall Non-Reference Sites . . . 6

1.9 Fraction of Individuals with Pathogenic Sites . . . 8

1.10 Common Pathogenic Variants by Ancestry . . . 10

2 Penetrance Estimates **11**

2.1 Bayes' Rule as a Model for Estimating Penetrance . . . 11

2.2 Collect and Aggregate Allele Frequencies at the Disease-Level . . . 11

2.3 Bootstrapped Distribution of Penetrance . . . 13

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)

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/

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

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: 43274 x 2516 (selected rows/columns):
```

	GENE	AF_1000G	VAR_ID	CHROM	POS	ID
62715	MYBPC3	0.000199681	11_47352958_G_A	11	47352958	rs527543611
62716	MYBPC3	0.000199681	11_47352974_C_T	11	47352974	rs541031071
62717	MYBPC3	0.000199681	11_47353028_C_T	11	47353028	rs564117422
62718	MYBPC3	0.018770000	11_47353058_C_T	11	47353058	rs11570121
62719	MYBPC3	0.000199681	11_47353134_C_T	11	47353134	rs549643481

Table continues below

	REF	ALT	HG00096	HG00097	HG00099	HG00100	HG00101	HG00102
62715	G	A	0	0	0	0	0	0
62716	C	T	0	0	0	0	0	0
62717	C	T	0	0	0	0	0	0
62718	C	T	0	0	0	0	0	0
62719	C	T	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: 31729 x 49 (selected rows/columns):
```

	GENE	AF_GNOMAD	AF_GNOMAD_NFE
22531	MYH7	0.000004119	0.000000000000
24995	TPM1	0.000032280	0.000000000000
6720	DSG2	0.000004063	0.000008958486
14845	LDLR	0.000004061	0.000000000000
541	PKP2	0.000004066	0.000000000000

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
NA19027	LWK	AFR	male
HG02028	KHV	EAS	female
HG01524	IBS	EUR	male
NA20514	TSI	EUR	female
HG00362	FIN	EUR	female
HG03788	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	42826
ACMG LP/P	9139
ACMG LP/P in gnomAD	662
ACMG LP/P in 1000 Genomes	53

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

Subset_gnomAD	Number_of_Variants
ACMG in gnomAD	31729
ClinVar-ACMG in gnomAD	4089
LP/P-ACMG in gnomAD	662

1.8 Overall Non-Reference Sites

1.8.0.1 For 1000 Genomes

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

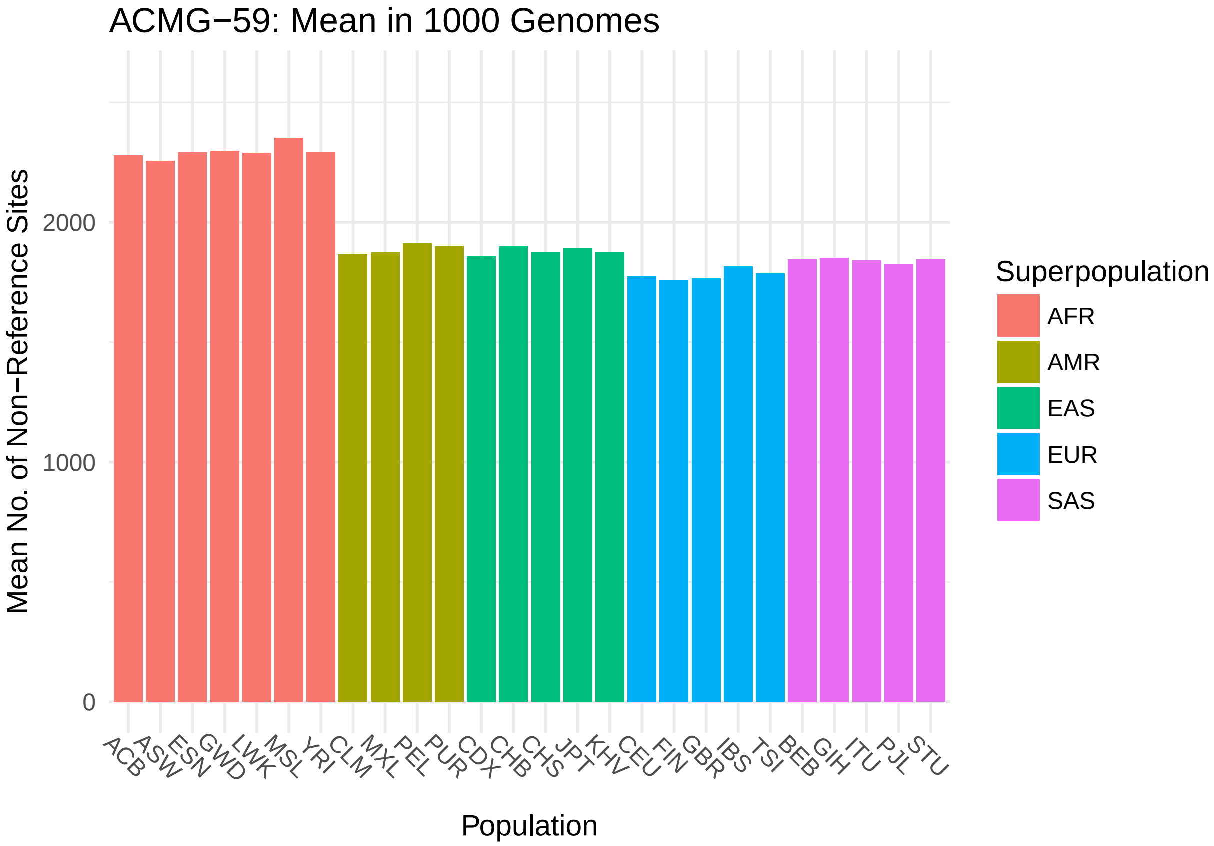
Ex: the genotype of 3 variants in 3 people looks like this:

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

Count the number of non-reference sites per individual:

HG00366	HG00367	HG00368
0	0	0

```
## Mean = 0
```



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

1.8.0.2 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:

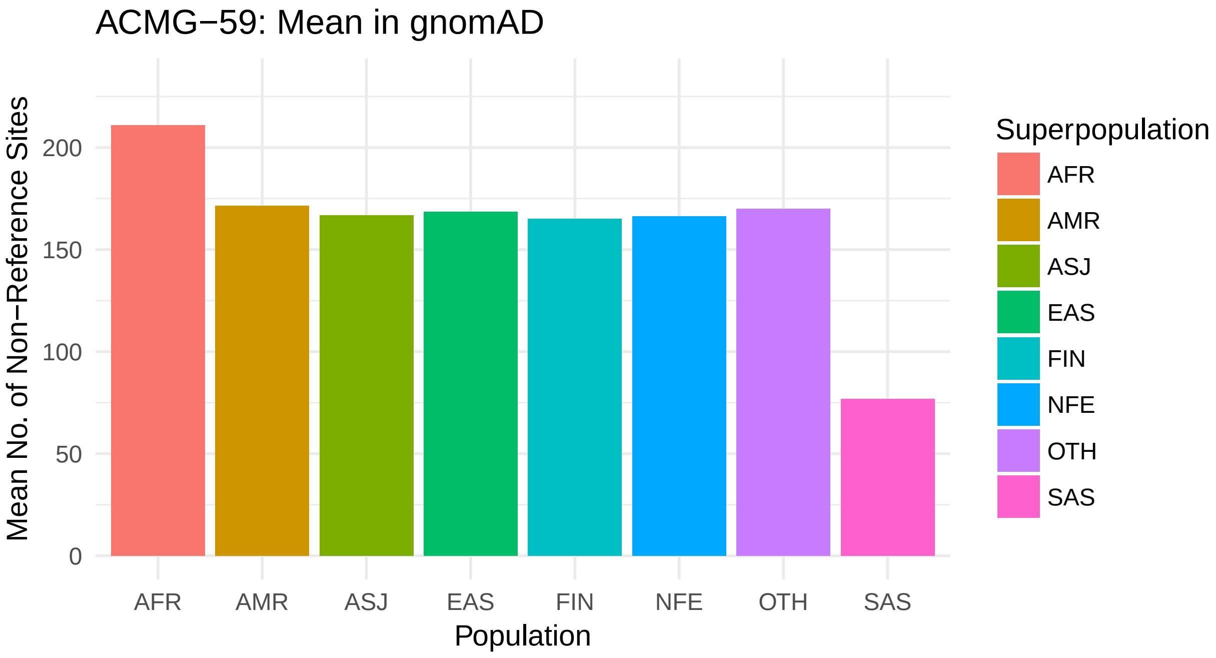
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

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



1.9 Fraction of Individuals with Pathogenic Sites

1.9.0.1 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.

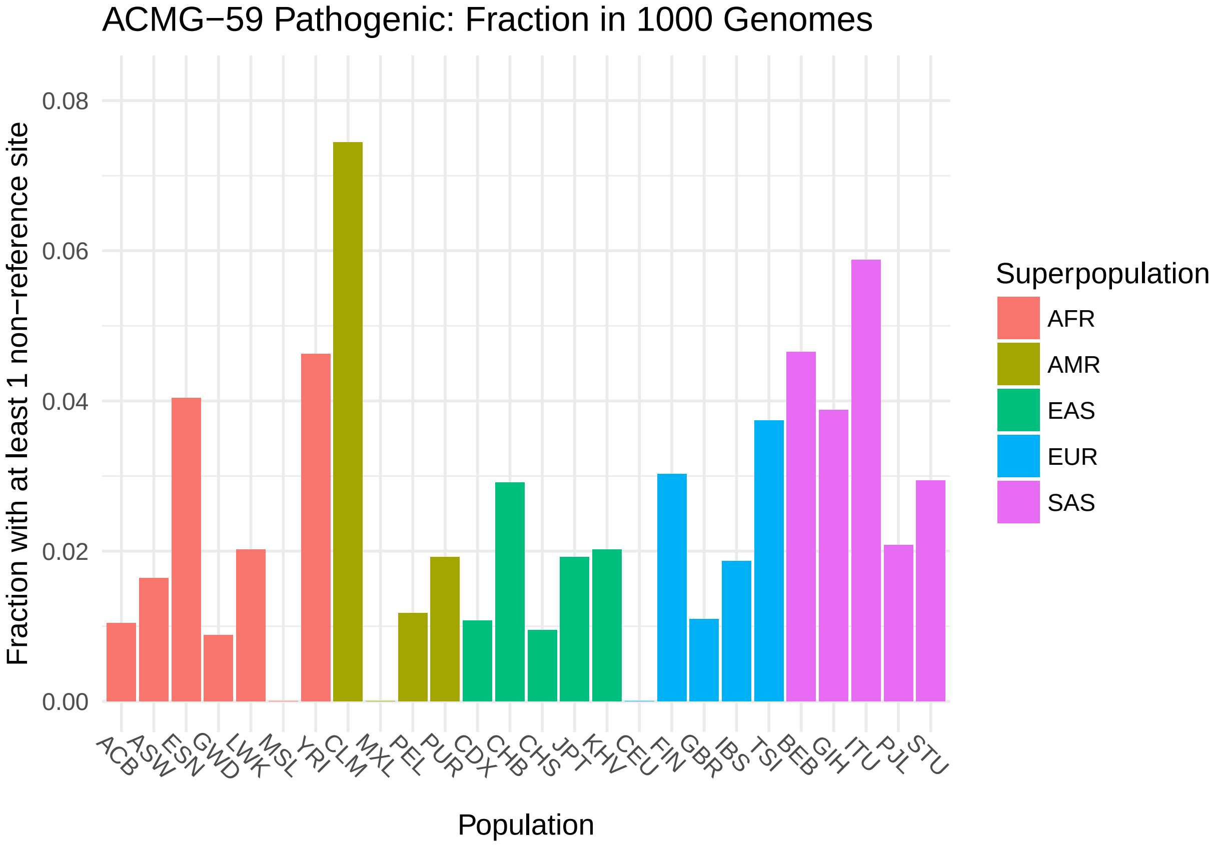
Ex: the genotype of 3 variants in 3 people looks like this:

	HG00366	HG00367	HG00368
Variant 1	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):

HG00366	HG00367	HG00368
0	0	0

```
## Mean = 0
```



1.9.0.2 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:

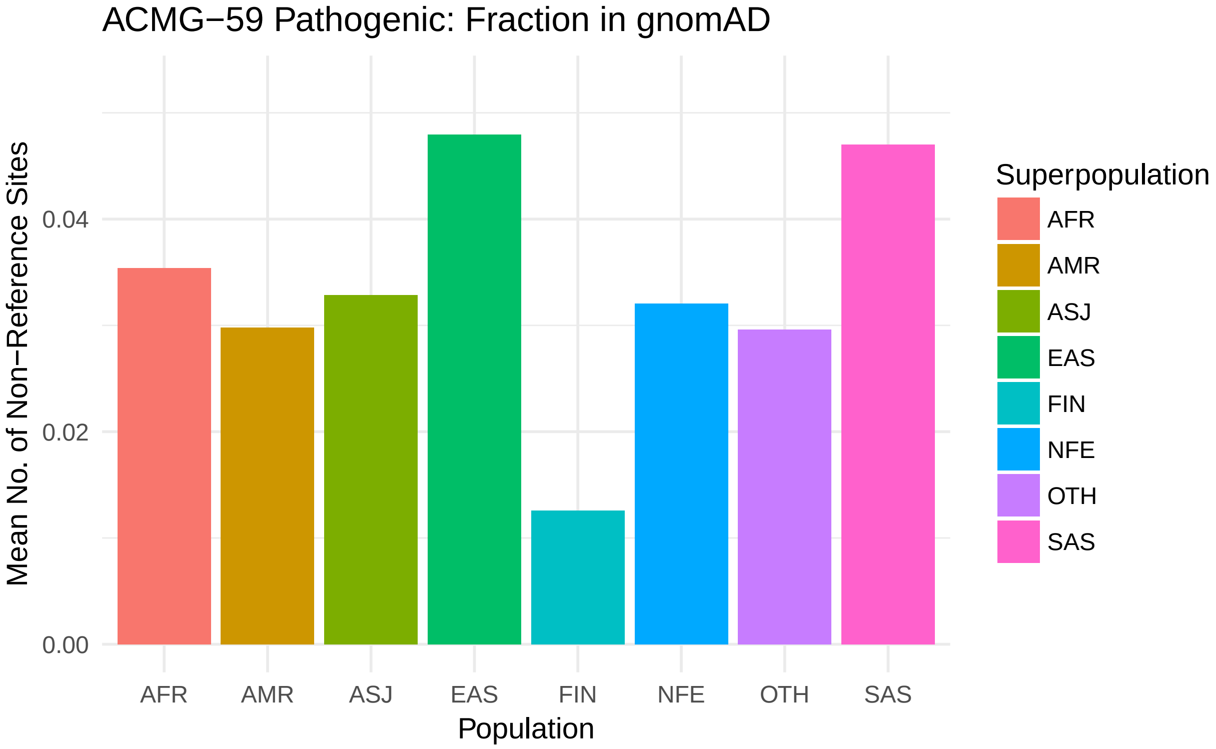
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

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:

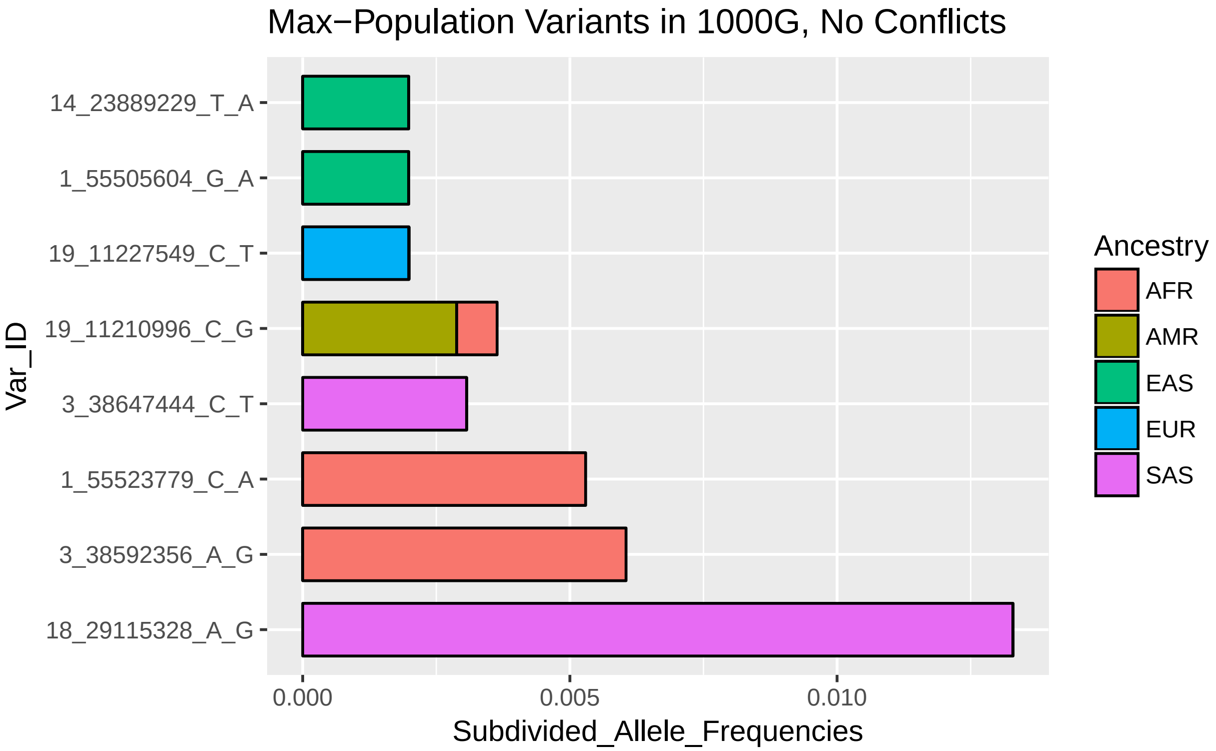
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.19	0.36	0	0	0.51
Variant 2	0.36	0	0.51	0	0.19

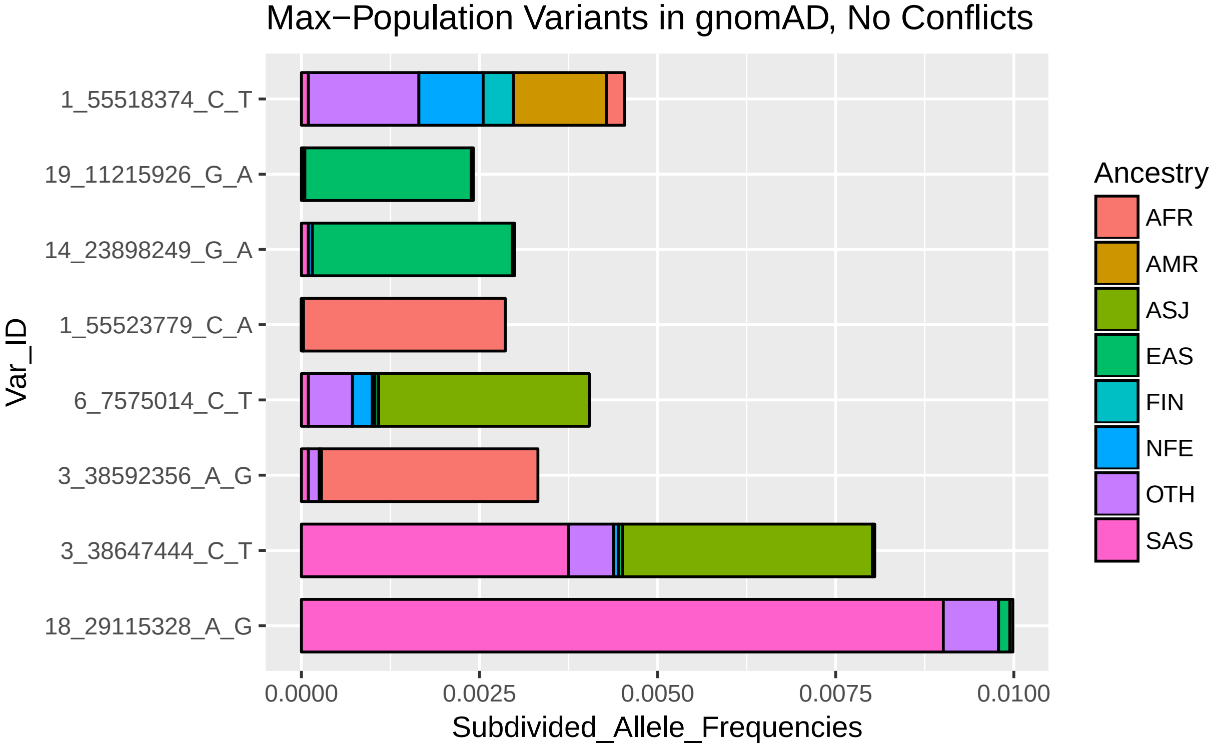
The expected (mean) number of non-reference sites is given by $1 - \prod(1 - AF)^2$.

AFR	AMR	EAS	EUR	SAS
0.4816	0.36	0.51	0	0.6031



1.10 Common Pathogenic Variants by Ancestry





2 Penetrance Estimates

2.1 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 $= P(D_x|V_x)$

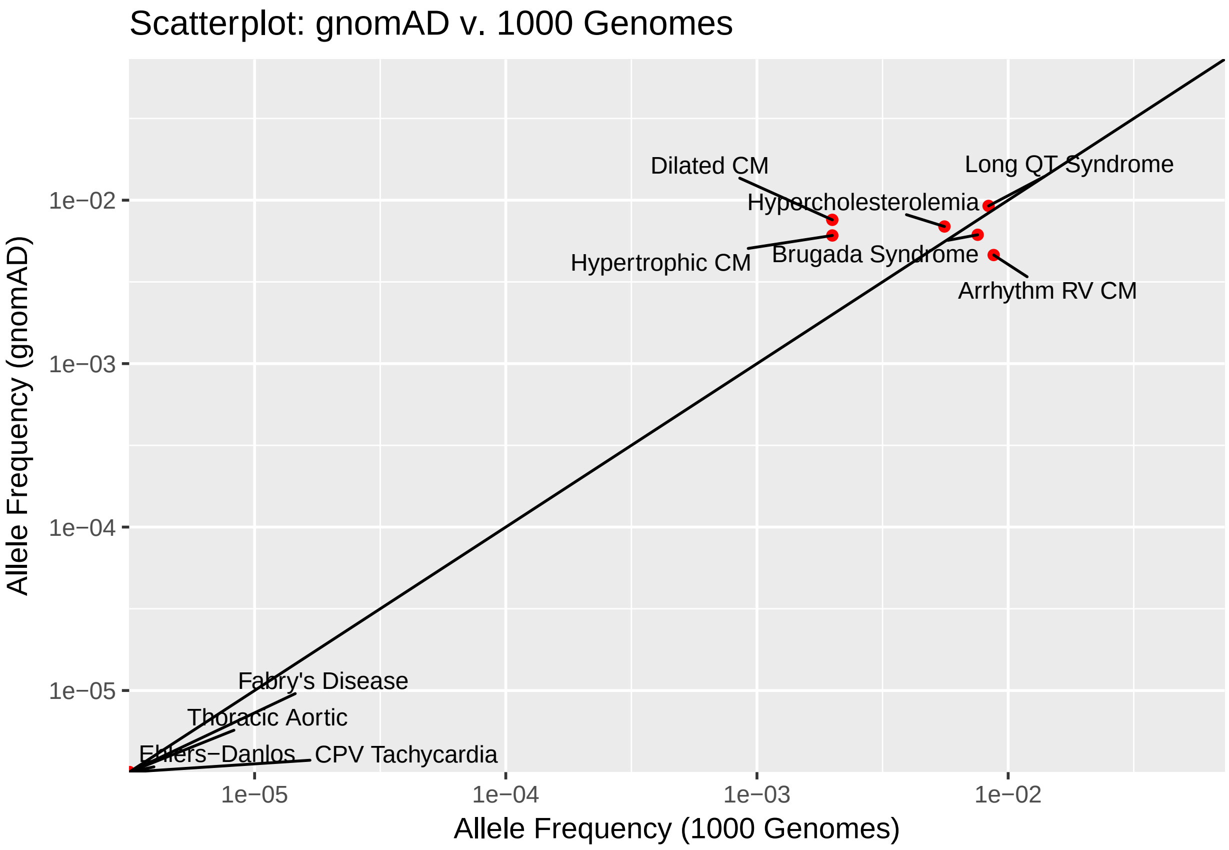
By Bayes' Rule, the penetrance of a variant related to disease x may be defined as:

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

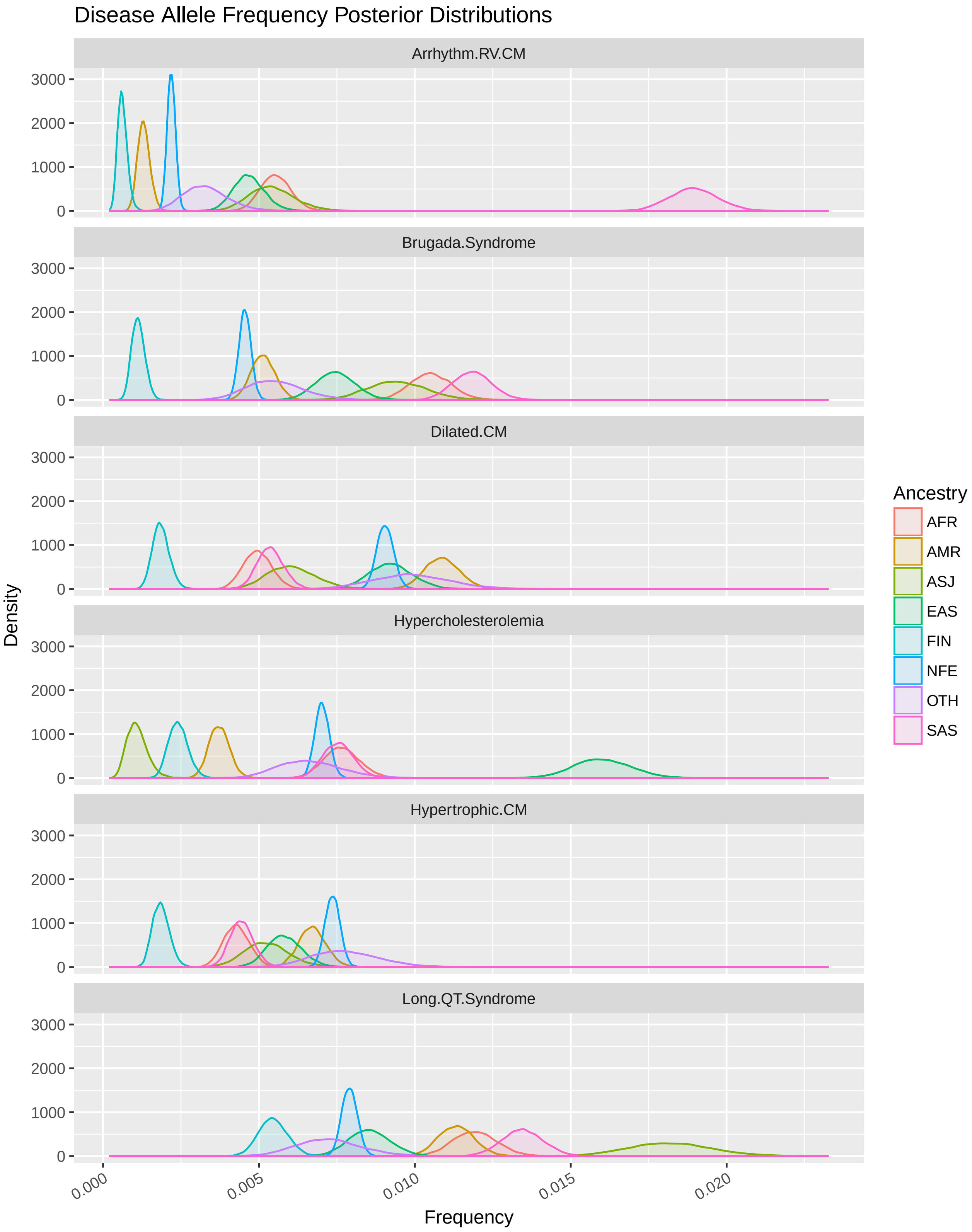
To compute penetrance estimates for each of the diseases related to the ACMG-59 genes, we will use the prevalence data we collected into `Literature_Prevalence_Estimates.csv`, allele frequency data from 1000 Genomes/ExAC/gnomAD, and a broad range of values for case allele frequency.

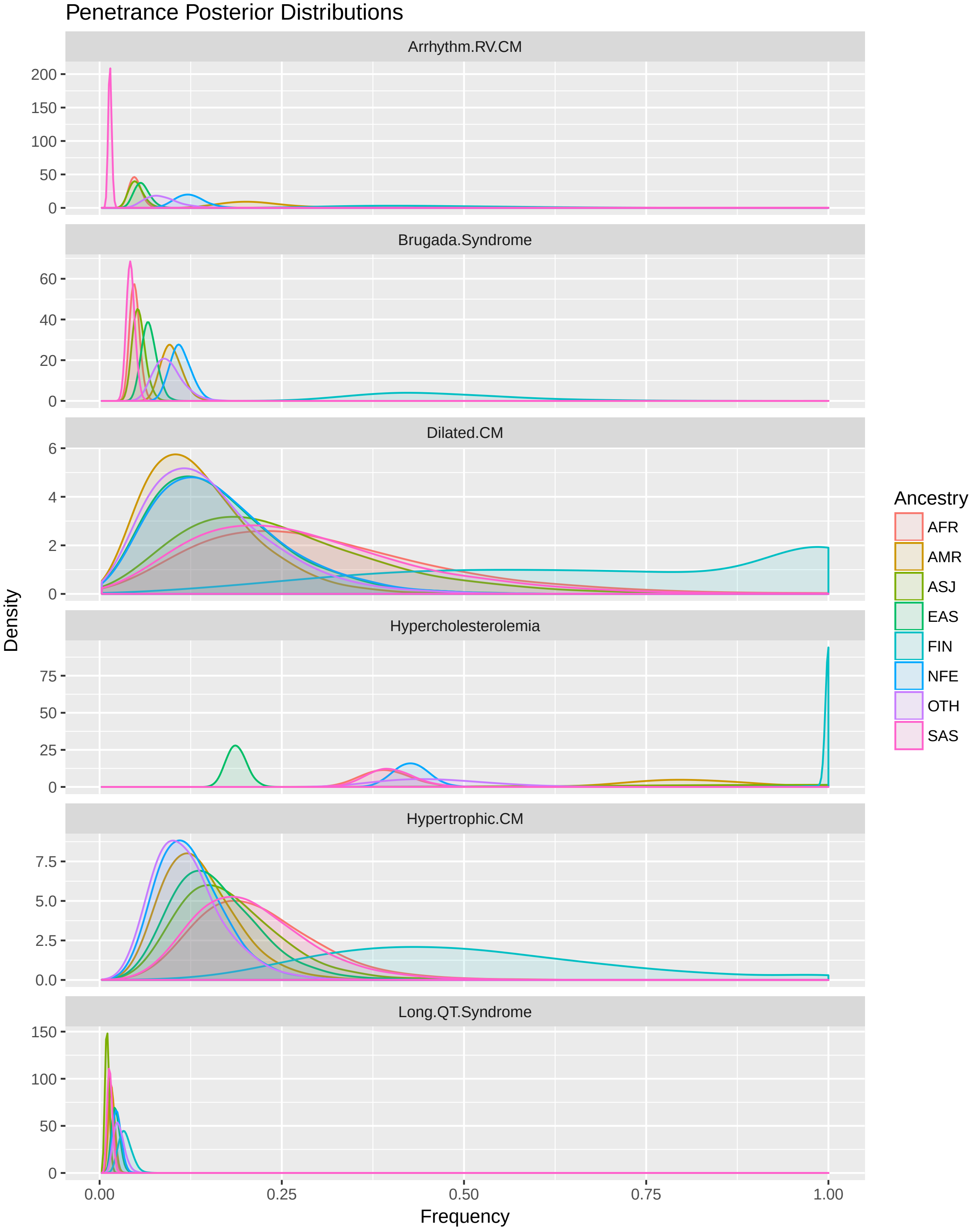
2.2 Collect and Aggregate Allele Frequencies at the Disease-Level

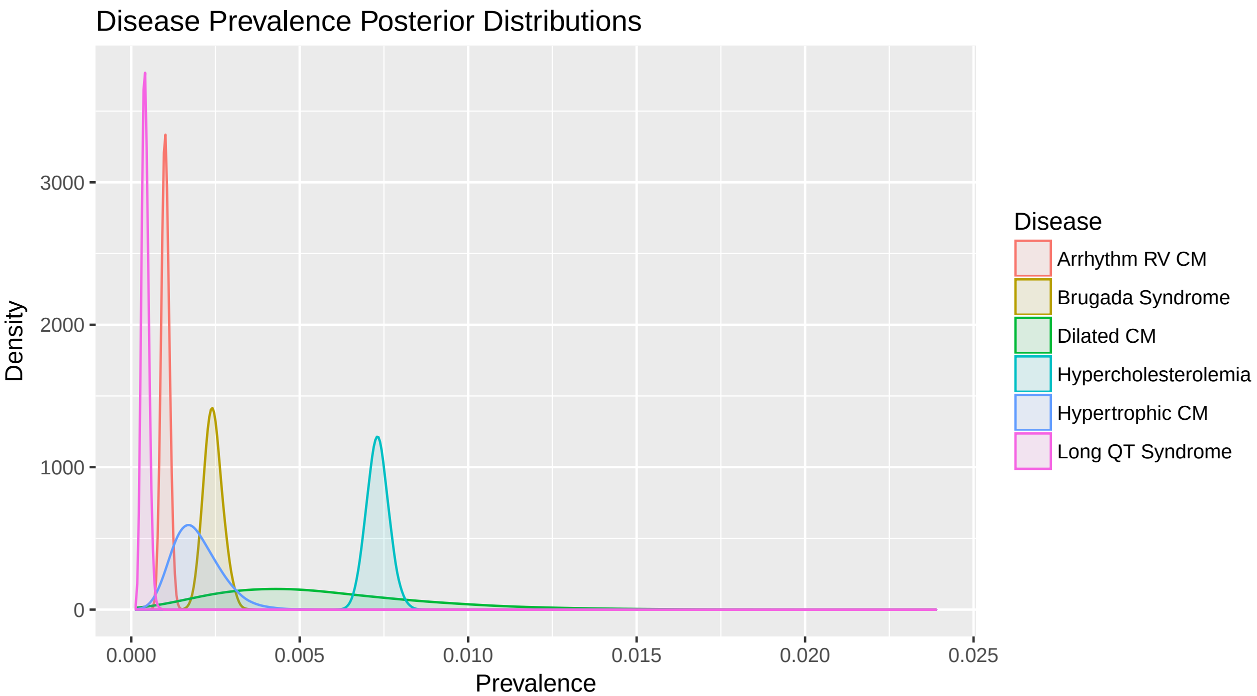
We define AF(disease) as the probability of having at least 1 variant associated with the disease. The variants can be assigned to diseases in two ways:
(1) By associating it by MIM. An MIM code is assigned for around 31% of assertions in each dataset.
(1) By associating it by MedGen. An MIM code is 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).

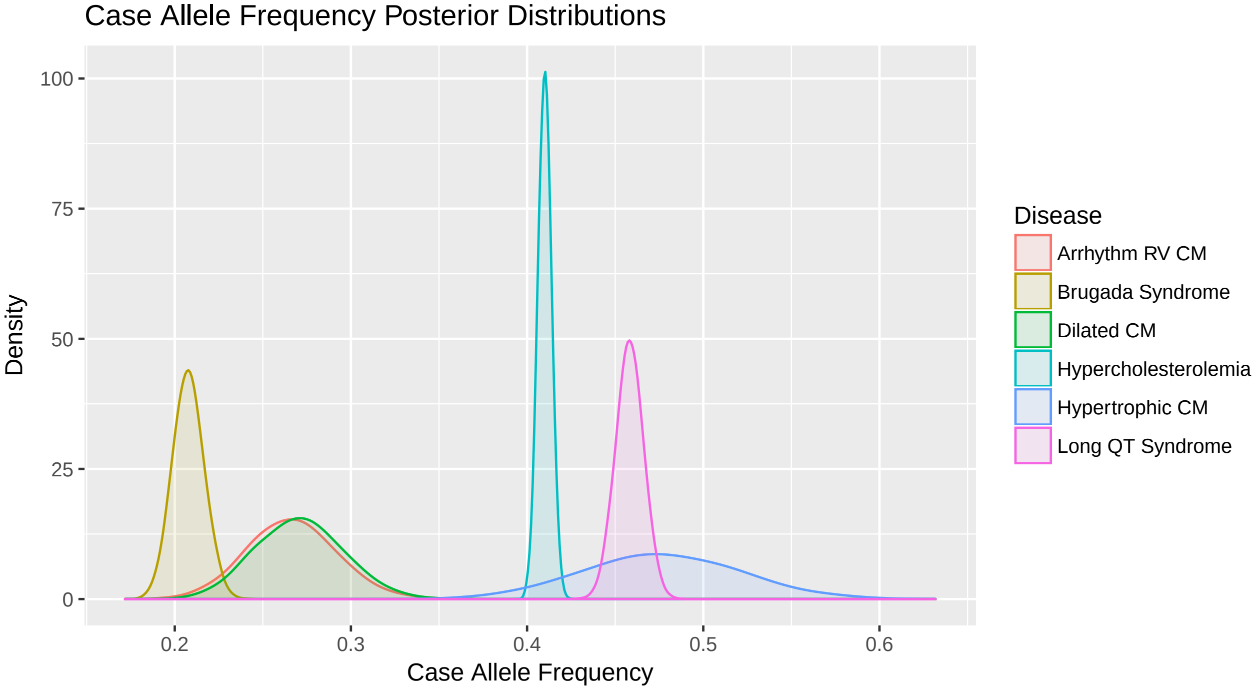

Scatterplot: gnomAD v. 1000 Genomes
Allele Frequency (gnomAD)
1e−02
1e−03
1e−04
1e−05
Allele Frequency (1000 Genomes)
1e−05
1e−04
1e−03
1e−02
Dilated CM
Long QT Syndrome
Hypercholesterolemia
Hypertrophic CM
Brugada Syndrome
Arrhythm RV CM
Fabry's Disease
Thoracic Aortic
Ehlers−Danlos
CPV Tachycardia

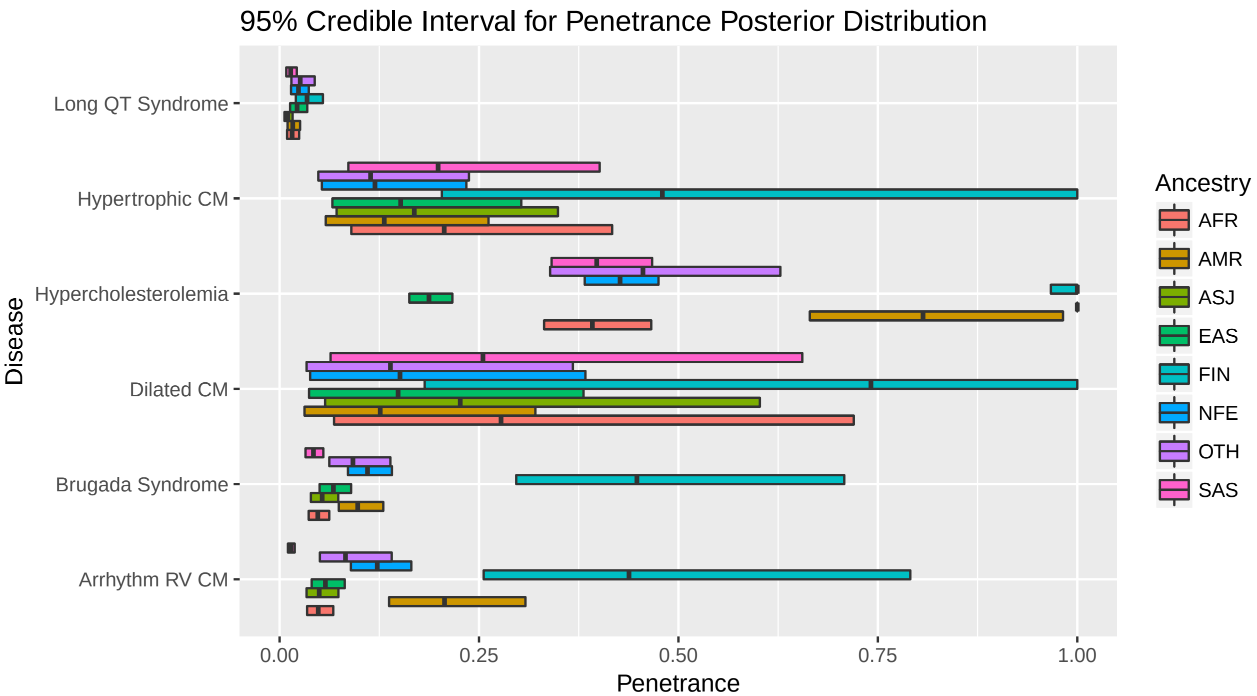
2.3 Bootstrapped Distribution of Penetrance

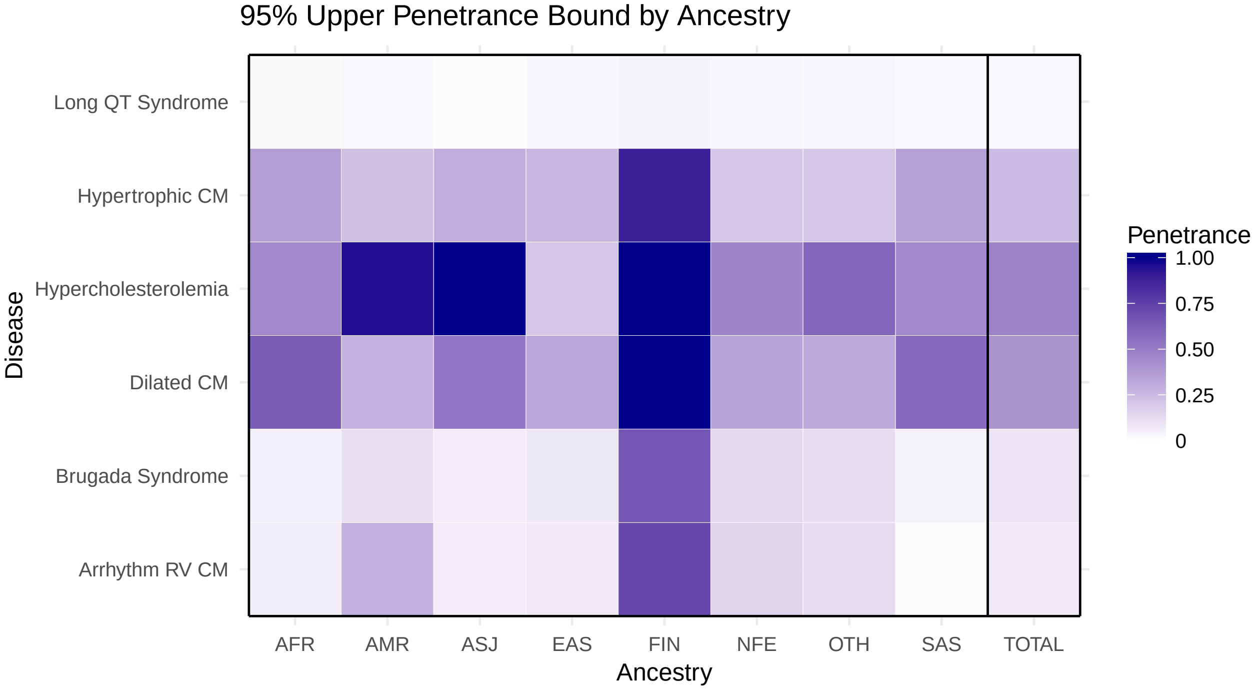



Penetrance Posterior Distributions
Arrhythm.RV.CM
200
150
100
50
0
Brugada.Syndrome
60
40
20
0
Dilated.CM
6
4
2
0
Hypercholesterolemia
75
50
25
0
Hypertrophic.CM
7.5
5.0
2.5
0.0
Long.QT.Syndrome
150
100
50
0
0.00
0.25
0.50
0.75
1.00
Frequency
Density
Ancestry
AFR
AMR
ASJ
EAS
FIN
NFE
OTH
SAS


Disease Prevalence Posterior Distributions
Density
3000
2000
1000
0
0.000
0.005
0.010
0.015
0.020
0.025
Prevalence
Disease
Arrhythm RV CM
Brugada Syndrome
Dilated CM
Hypercholesterolemia
Hypertrophic CM
Long QT Syndrome


Case Allele Frequency Posterior Distributions
Density
100
75
50
25
0
0.2
0.3
0.4
0.5
0.6
Case Allele Frequency
Disease
Arrhythm RV CM
Brugada Syndrome
Dilated CM
Hypercholesterolemia
Hypertrophic CM
Long QT Syndrome


95% Credible Interval for Penetrance Posterior Distribution
Disease
Long QT Syndrome
Hypertrophic CM
Hypercholesterolemia
Dilated CM
Brugada Syndrome
Arrhythm RV CM
0.00
0.25
0.50
0.75
1.00
Penetrance
Ancestry
AFR
AMR
ASJ
EAS
FIN
NFE
OTH
SAS


95% Upper Penetrance Bound by Ancestry
Disease
Long QT Syndrome
Hypertrophic CM
Hypercholesterolemia
Dilated CM
Brugada Syndrome
Arrhythm RV CM
AFR
AMR
ASJ
EAS
FIN
NFE
OTH
SAS
TOTAL
Ancestry
Penetrance
1.00
0.75
0.50
0.25
0