

ACMG-ClinVar Penetrance Markdown

James Diao, under the supervision of Arjun Manrai

October 28, 2016

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	Collect 1000 Genomes Phase 3 Populations Map	4
1.5	Import and Process 1000 Genomes VCFs	5
1.6	Import and Process ExAC VCFs	5
1.7	Merge ClinVar with 1000 Genomes and ExAC	6
1.8	Comparison with ClinVar Browser Query Results	7
2	Plot Summary Statistics Across Populations	8
2.1	Overall Non-Reference Sites	8
2.2	Pathogenic Non-Reference Sites	10
2.3	Fraction of Individuals with Pathogenic Sites	11
2.4	Test Statistics for Ancestral Differences	13
2.5	Common Pathogenic Variants by Ancestry	14
3	Penetrance Estimates	15
3.1	Import Literature-Based Disease Prevalence Data	15
3.2	Collect and Aggregate Allele Frequencies at the Disease-Level	15
3.3	Comparisons of AF(disease) by Dataset and Collection Methods	16
3.4	Penetrance as a Function of $P(V D)$	17
3.5	Penetrance as a Function of Prevalence	18

Working Directory: /Users/jamesdiao/Documents/Kohane_Lab/2016-paper-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 64 x 4 (selected rows):

Disease_Name	Disease_MIM	Gene_Name	Gene_MIM
Adenomatous polyposis coli	175100	APC	611731
Breast-ovarian cancer, familial 1	604370	BRCA1	113705
Brugada syndrome 1	601144	SCN5A	600163
Dilated cardiomyopathy 1A	115200	LMNA	150330
Familial hypercholesterolemia	143890	APOB	107730
Familial hypertrophic cardiomyopathy 1	192600	MYH7	160760
Retinoblastoma	180200	RB1	614041

ACMG-56 Genes:

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

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 117420 x 14 (selected rows/columns):

VAR_ID	CHROM	POS	ID	REF	ALT	CLNSIG
1_955597_G_T	1	955597	rs115173026	G	T	2
1_955619_G_C	1	955619	rs201073369	G	C	255
1_957605_G_A	1	957605	rs756623659	G	A	5

Table continues below

CLNDBN	CLNDSDBID	INTERP
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: 56 x 6 (selected rows):

gene	name	chrom	start	end	downloaded
APC	NM_001127511	5	1.12e+08	112181936	TRUE
MYH11	NM_001040113	16	15796991	15950887	TRUE
ACTA2	NM_001141945	10	90694830	90751154	TRUE
MYLK	NM_001321309	3	123331142	123603149	TRUE
TMEM43	NM_024334	3	14166439	14185180	TRUE

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

1.4 Collect 1000 Genomes Phase 3 Populations Map

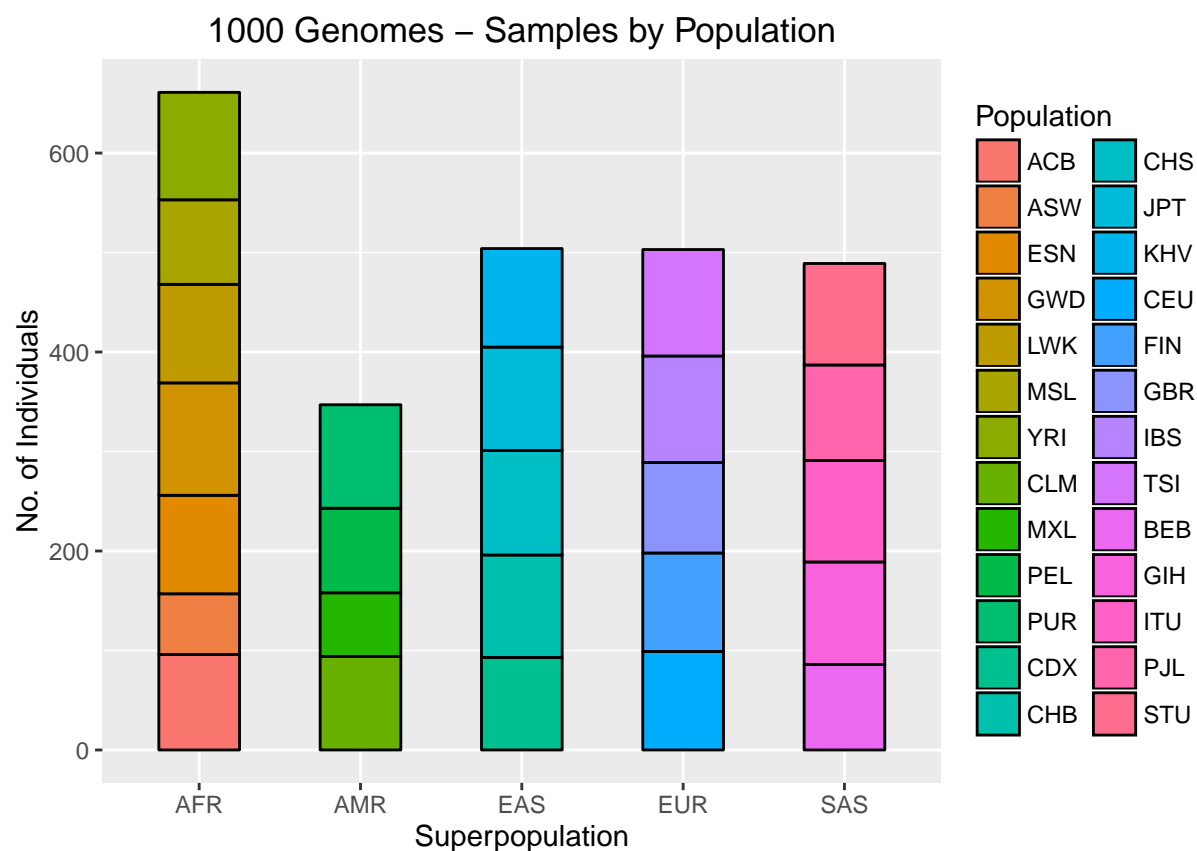
This will allow us to assign genotypes from the 1000 Genomes VCF to ancestral groups.

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
NA19121	YRI	AFR	male
HG02813	GWD	AFR	male
NA19036	LWK	AFR	female
NA19037	LWK	AFR	female
HG02339	ACB	AFR	female
HG01323	PUR	AMR	female
HG00533	CHS	EAS	male
HG00133	GBR	EUR	female
NA20854	GIH	SAS	female
HG03016	PJL	SAS	female

Population Distribution



1.5 Import and Process 1000 Genomes VCFs

- Unnest the data frames to 1 row per variant_ID key (CHROM_POSITION_REF_ALT).
- Remove all insertions, deletions, CNV, etc, and keep only missense variants (1 REF, 1 ALT)
- 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: 139335 x 2516 (selected rows/columns):

GENE	AF_1000G	VAR_ID	CHROM	POS	ID	REF	ALT
APC	0.0001997	5_112043211_A_G	5	1.12e+08	rs554351451	A	G
APC	0.0001997	5_112043231_G_A	5	1.12e+08	rs575784409	G	A
APC	0.005391	5_112043234_C_T	5	1.12e+08	rs115658307	C	T
APC	0.0001997	5_112043252_G_A	5	1.12e+08	rs558562104	G	A
APC	0.008786	5_112043263_C_T	5	1.12e+08	rs138386816	C	T

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
0	0	0	0	0	0
0	0	0	0	0	0

1.6 Import and Process ExAC VCFs

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

Processed ExAC VCFs: 58873 x 45 (selected rows/columns):

GENE	AF_EXAC	AF_EXAC_AFR	AF_EXAC_AMR	AF_EXAC_EAS	AF_EXAC_EUR
APC	8.13e-05	0	0	0	0
APC	8.131e-05	0	0	0	0
APC	0.1112	0.07979	0.1022	0	0.1063
APC	8.131e-05	0	0	0	0
APC	8.134e-05	0	0	0	0

Table continues below

AF_EXAC_SAS	VAR_ID	CHROM	POS	ID	REF	ALT
0.0001313	5_112043365_G_C	5	1.12e+08	.	G	C
0.0001313	5_112043382_A_G	5	1.12e+08	.	A	G
0.1185	5_112043384_T_G	5	1.12e+08	rs78429131	T	G
0.0001313	5_112043392_C_T	5	1.12e+08	.	C	T
0.0001313	5_112043412_C_G	5	1.12e+08	.	C	G

1.7 Merge ClinVar with 1000 Genomes and ExAC

Breakdown of ClinVar Variants

Subset_ClinVar	Number_of_Variants
Total ClinVar	117420
LP/P-ClinVar	33633
LP/P-ClinVar & ACMG	6971
LP/P-ClinVar & ACMG & ExAC	964
LP/P-ClinVar & ACMG & 1000 Genomes	147

Breakdown of ACMG-1000 Genomes Variants

Subset_1000_Genomes	Number_of_Variants
Total 1000_Genomes & ACMG	139335
1000_Genomes & ACMG & ClinVar	4339
1000_Genomes & ACMG & LP/P-ClinVar	147

Breakdown of ACMG-ExAC Variants

Subset_ExAC	Number_of_Variants
Total ExAC & ACMG	58873
ExAC & ACMG & ClinVar	9347
ExAC & ACMG & LP/P-ClinVar	964

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	508
In LP/P-ClinVar VCF	504
^ & ACMG & ExAC	48
^ & 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 Overall Non-Reference Sites

2.1.0.1 For 1000 Genomes

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

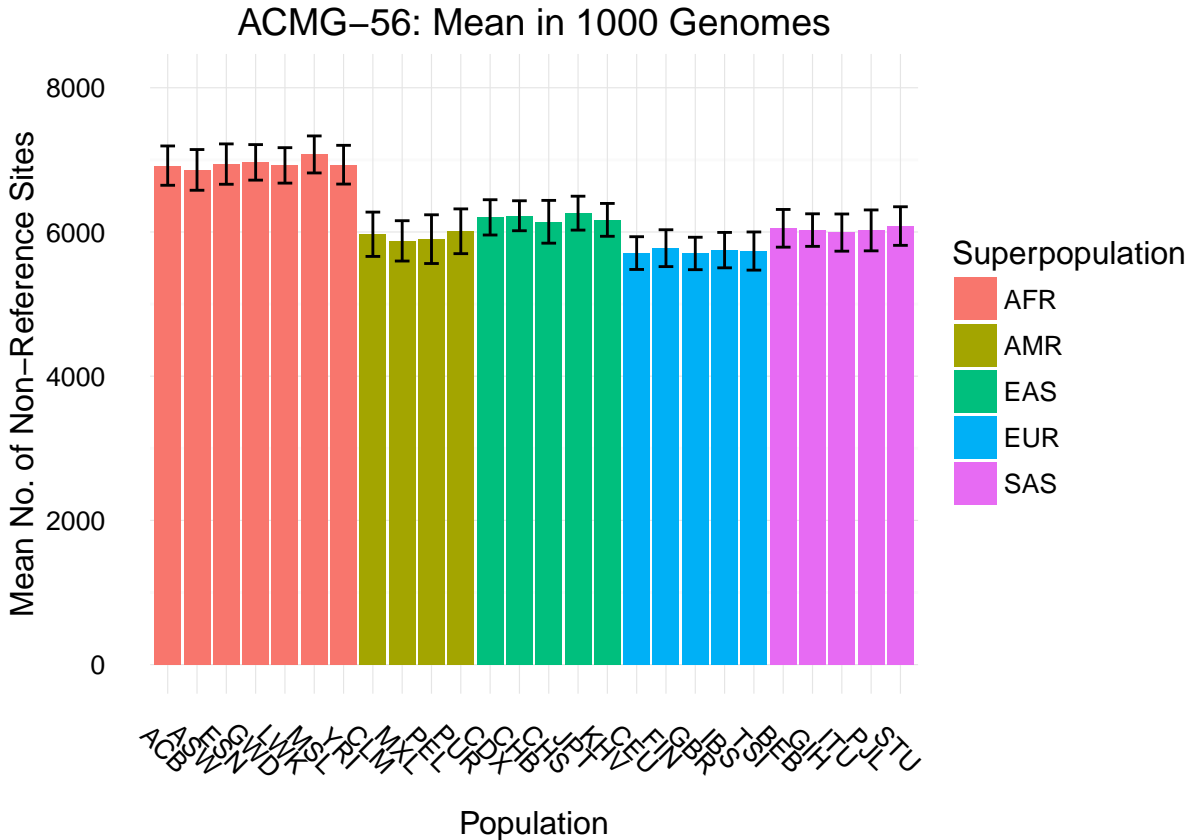
Ex: the genotype of 3 variants in 3 people looks like this:

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

Count the number of non-reference sites per individual:

HG00097	HG00099	HG00100
0	1	3

Mean = 1.33



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

2.1.0.2 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:

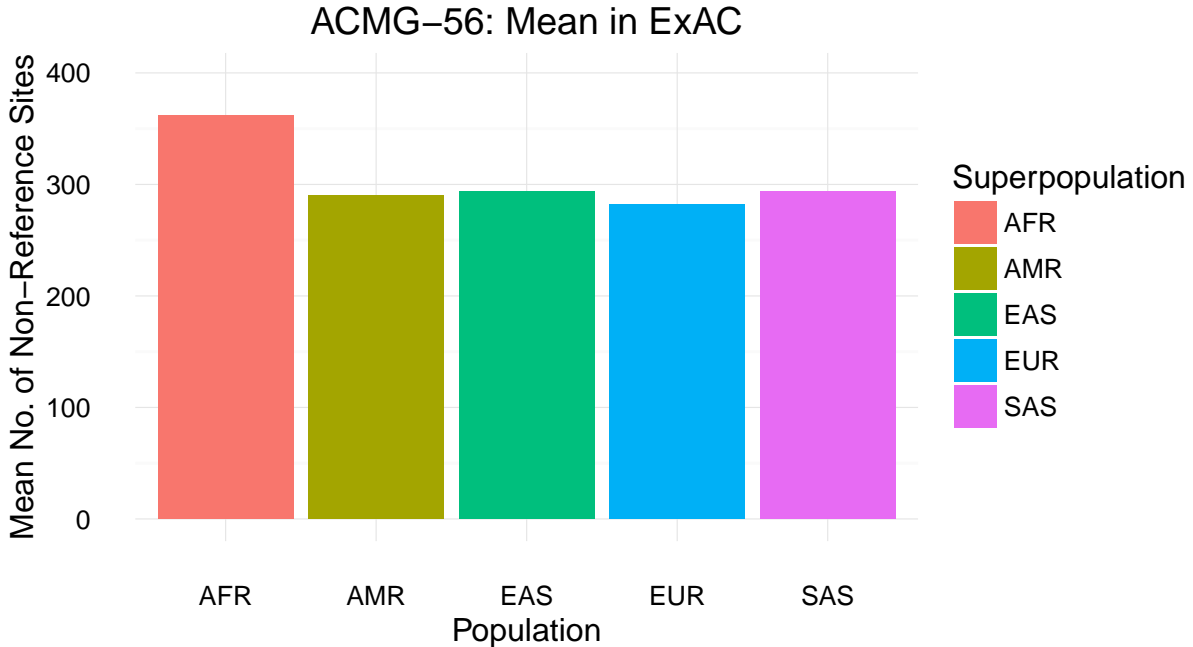
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

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 $\sum E(V_i) = \sum(\text{columns})$.

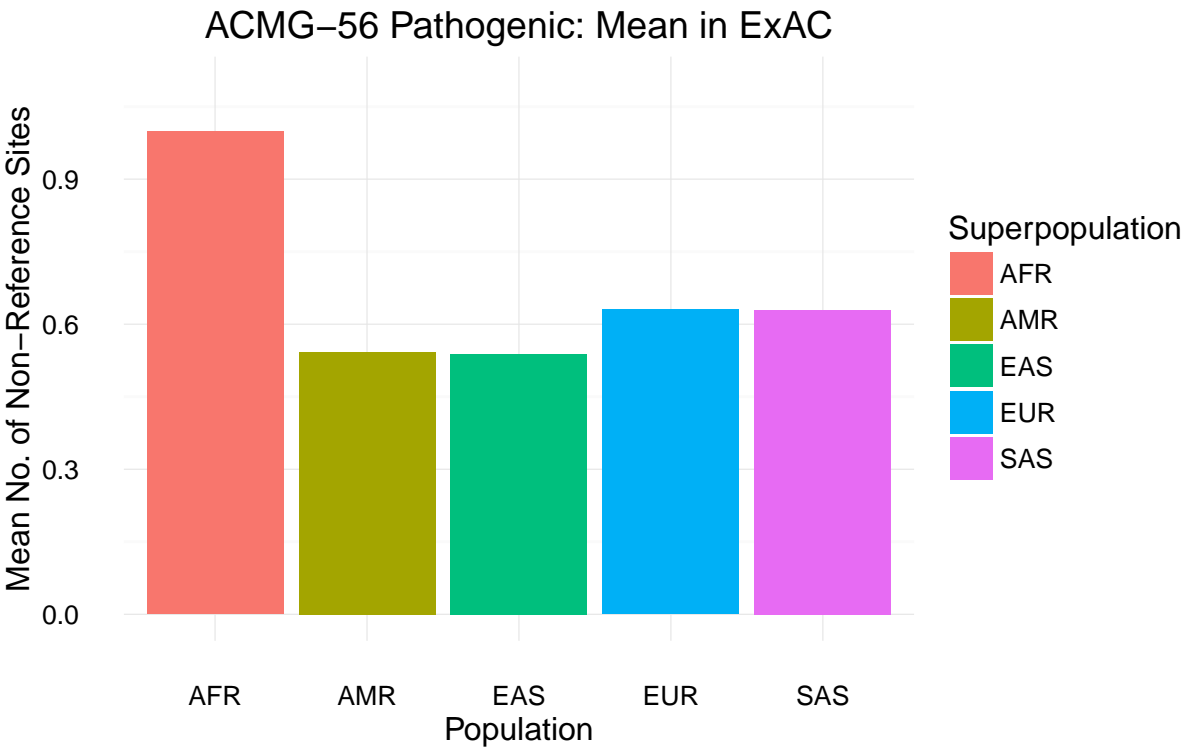
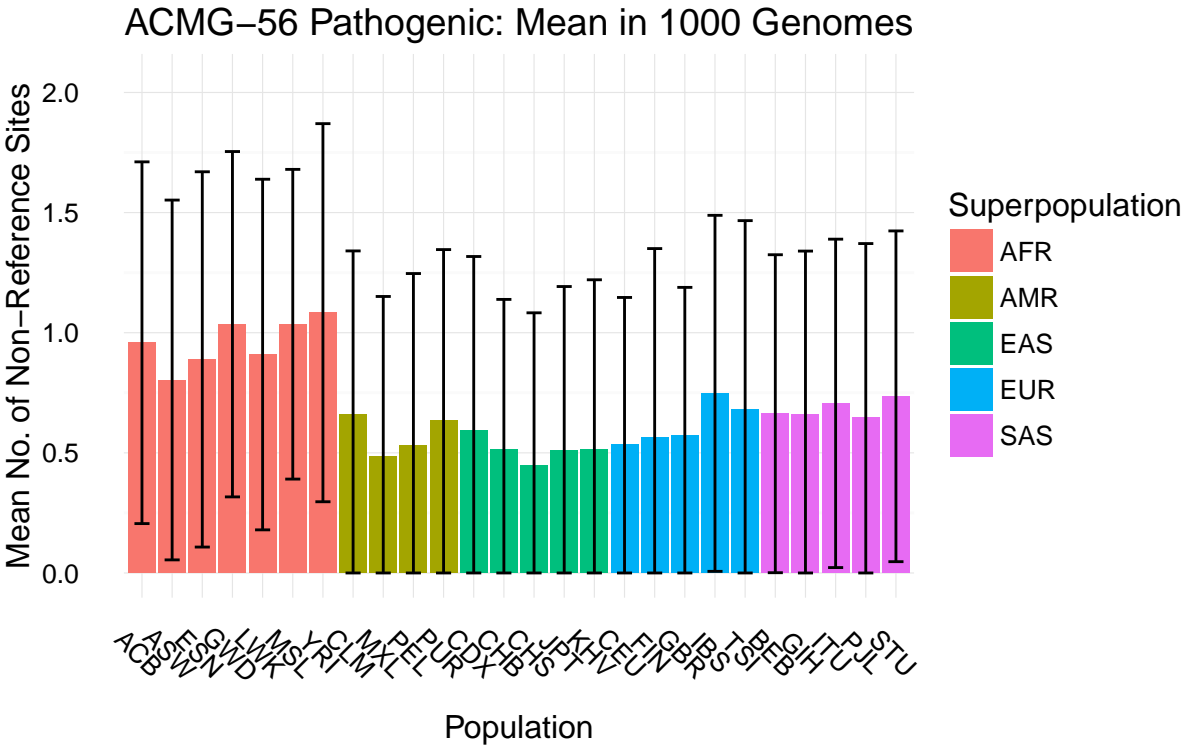
AFR	AMR	EAS	EUR	SAS
0.55	0.36	0.51	0	0.7



2.2 Pathogenic Non-Reference Sites

2.2.0.1 For 1000 Genomes and ExAC

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



2.3 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-56 genes.

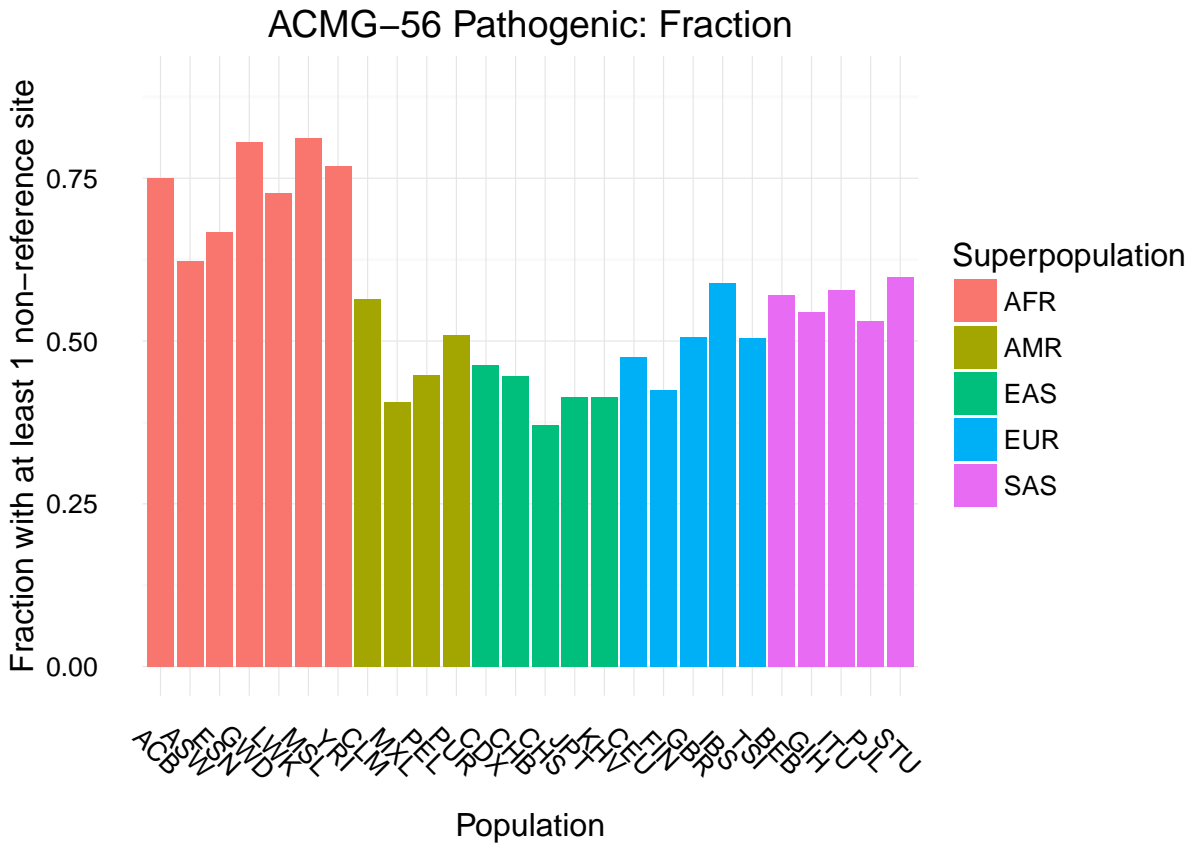
Ex: the genotype of 3 variants in 3 people looks like this:

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

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

HG00097	HG00099	HG00100
0	1	1

Mean = 0.667



2.3.0.2 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:

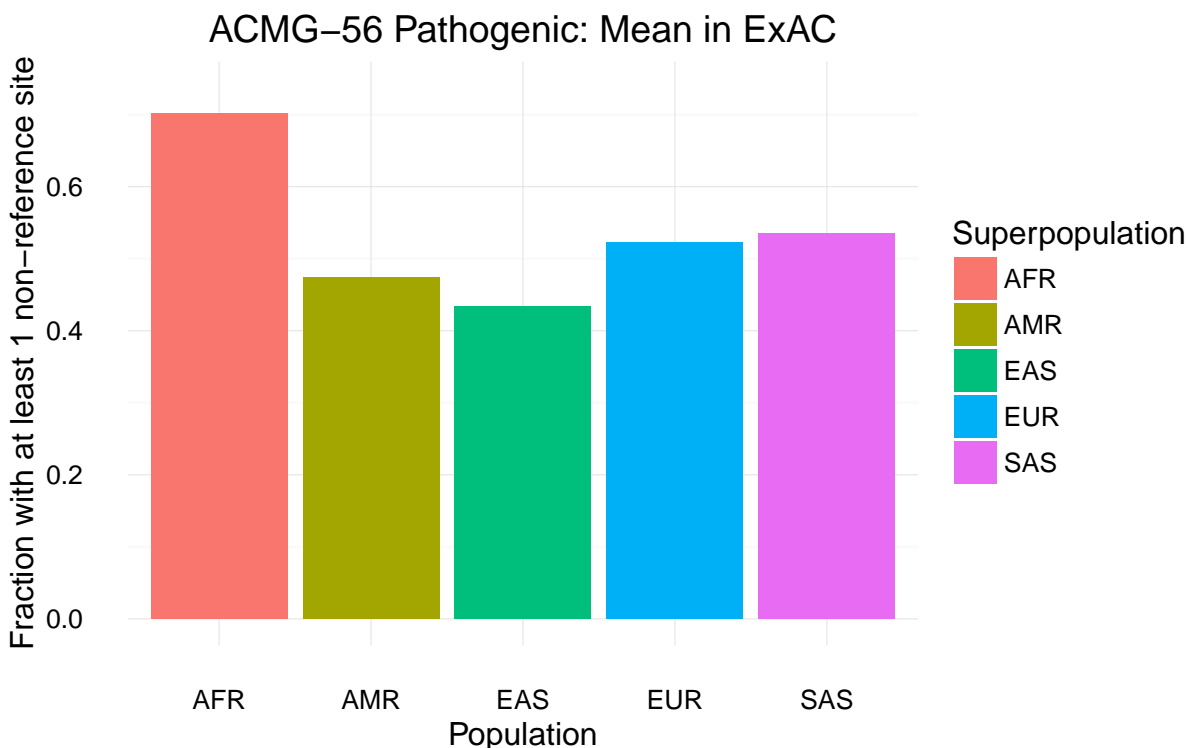
	AFR	AMR	EAS	EUR	SAS
Variant 1	0.1	0.2	0	0	0.3
Variant 2	0.2	0	0.3	0	0.1

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

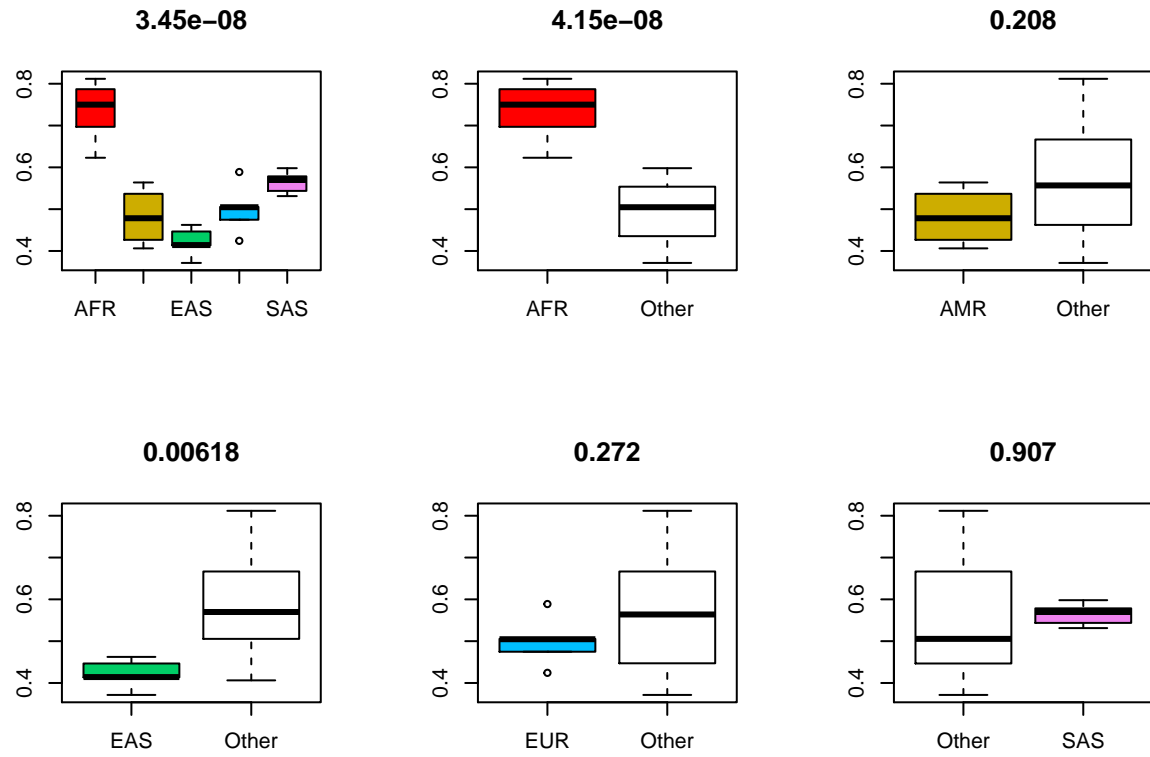
The expected (mean) number of non-reference sites is given by $1 - \prod (1 - AF)^2$.

AFR	AMR	EAS	EUR	SAS
0.4816	0.36	0.51	0	0.6031



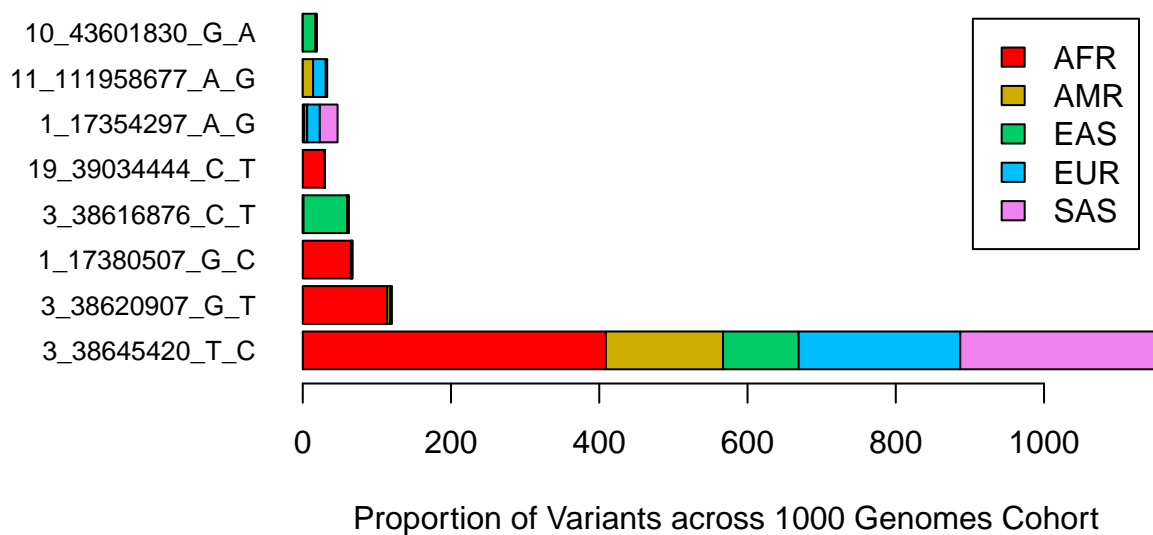
2.4 Test Statistics for Ancestral Differences

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.

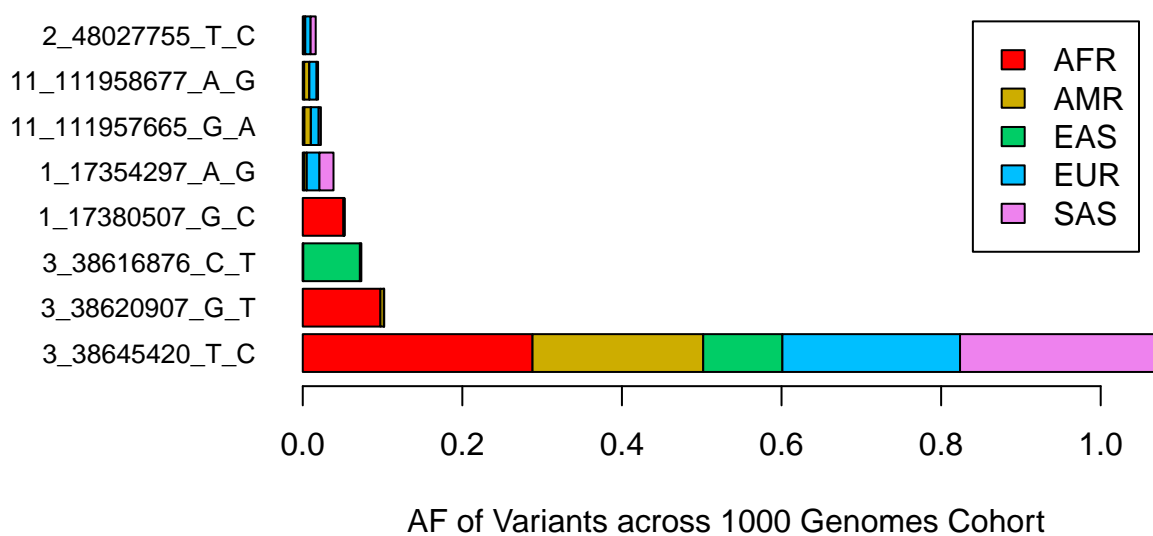


2.5 Common Pathogenic Variants by Ancestry

In 1000 Genomes



In ExAC



3 Penetrance Estimates

3.1 Import Literature-Based Disease Prevalence Data

Table of Literature-Based Estimates of Disease Prevalence 30 x 16 (selected rows/columns):

Gene	Disease	Disease_MIM	Tags
BRCA1;BRCA2	Breast-ovarian cancer familial	604370;612555	breast;ovarian
SCN5A	Brugada syndrome	601144	brugada
COL3A1	Ehlers-Danlos syndrome	130050	ehler;danlos
TP53	Li-Fraumeni syndrome	151623	fraumeni

Table continues below

Inverse.Prevalence.1	Inverse.Prevalence.2	year	first.author	citations
104	NA	2013	NA	NA
10000	2000	2006	Antzelevitch	11
20000	NA	2010	Malfait	116
20000	5000	1999	Schneider	47

3.2 Collect and Aggregate Allele Frequencies at the Disease-Level

We define $AF(\text{disease})$ as the probability of having at least 1 variant associated with the disease.

There are three ways of identifying which disease a variant belongs to:

- (1) By gene(s)
- (2) By disease MIM(s)
- (3) By keyword(s)

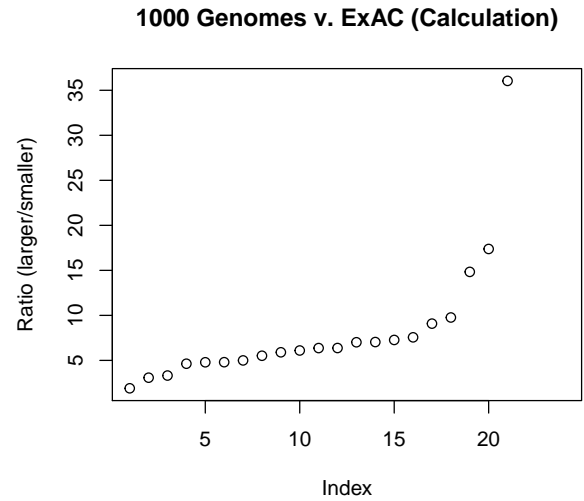
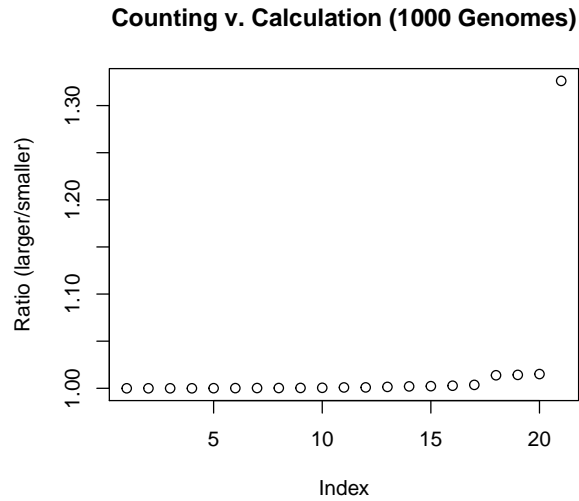
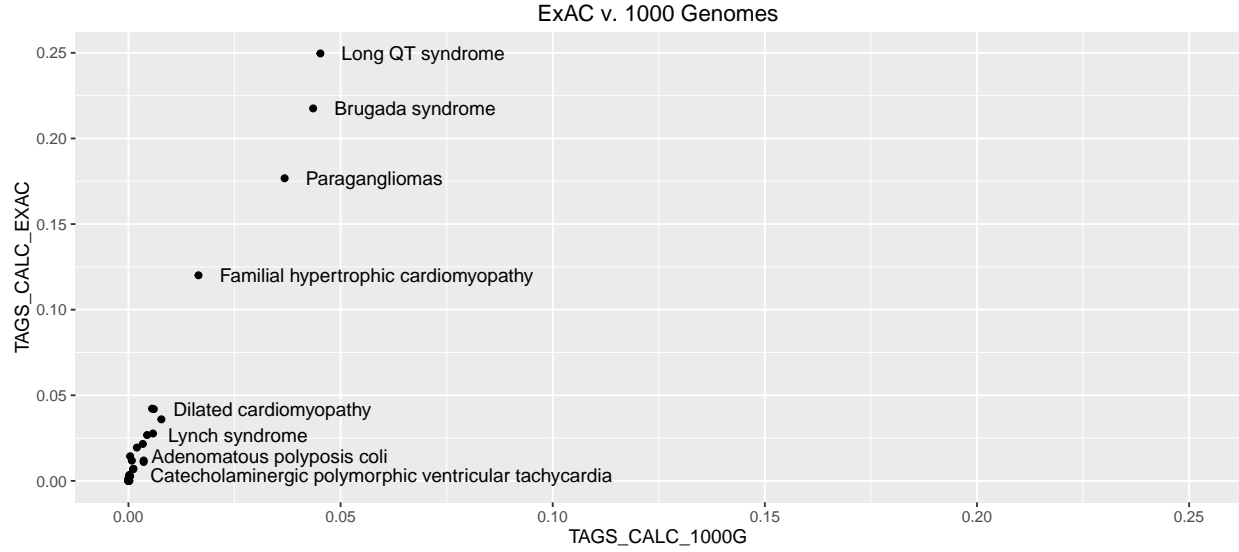
The frequencies across the relevant variants can be aggregated in two ways:

- (1) By direct counting, from genotype data in 1000 Genomes.
- (2) $AF(\text{disease}) = 1 - \prod_{\text{variant}} (1 - AF_{\text{variant}})$, from population data in ExAC (assumes independence).

3.3 Comparisons of AF(disease) by Dataset and Collection Methods

Correlation Table:

	TAGS_COUNT_1000G	TAGS_CALC_1000G	TAGS_CALC_EXAC
TAGS_COUNT_1000G	1	1	0.9908
TAGS_CALC_1000G	1	1	0.991
TAGS_CALC_EXAC	0.9908	0.991	1

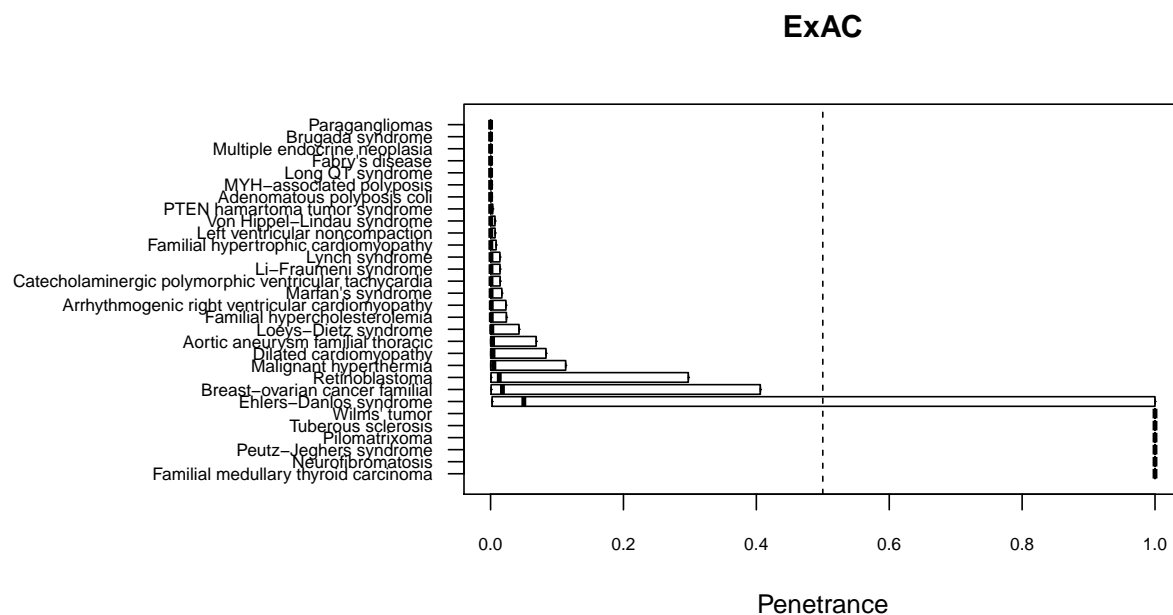
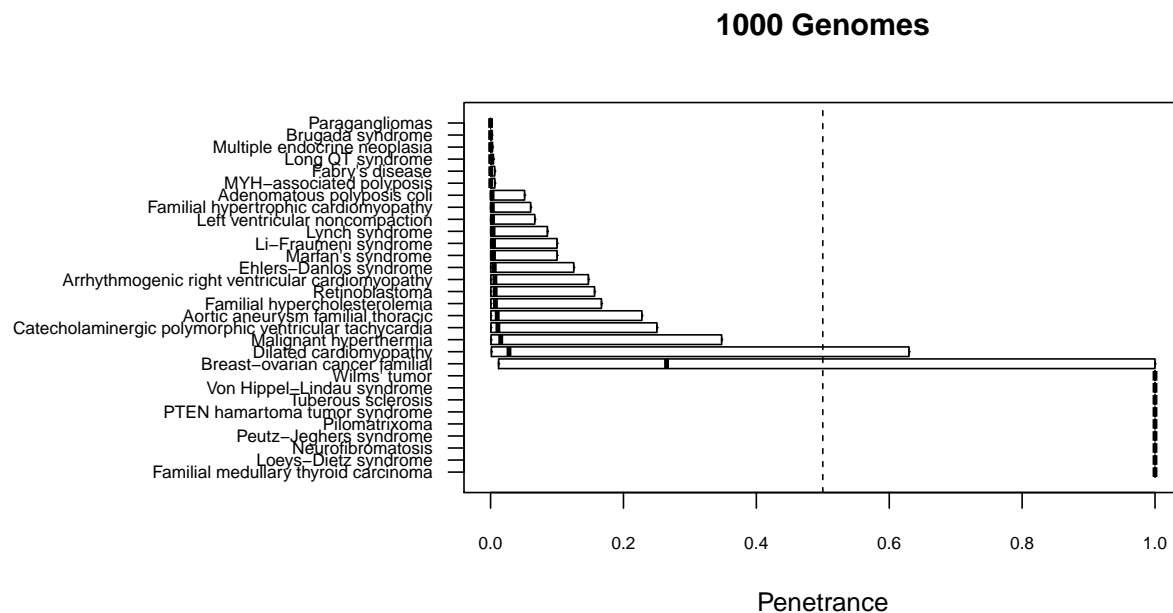


The median AF(disease) ratio between counting and calculation is: 1.001.

The median AF(disease) ratio between ExAC and 1000 Genomes is: 6.37.

3.4 Penetrance as a Function of $P(V|D)$

The left end of the boxplot indicates $P(V|D) = 0.001$,
the bold line in the middle indicates $P(V|D) = 0.02$,
the right end of the boxplot indicates $P(V|D) = 0.5$.

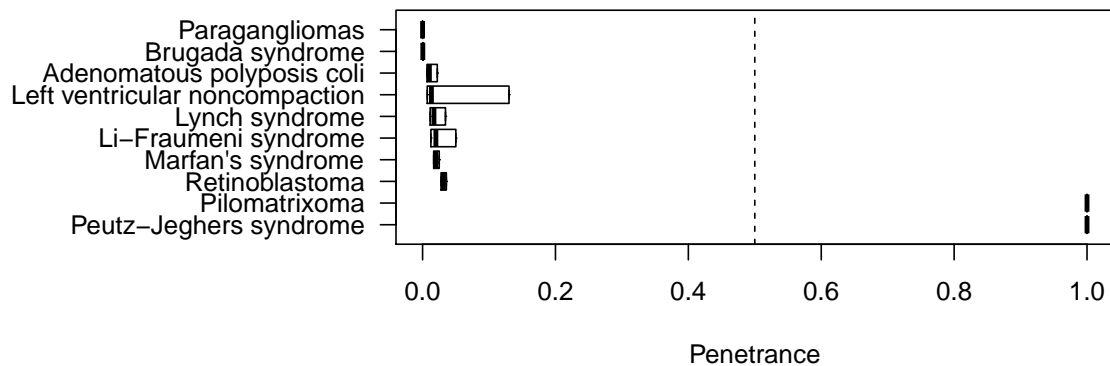


3.5 Penetrance as a Function of Prevalence

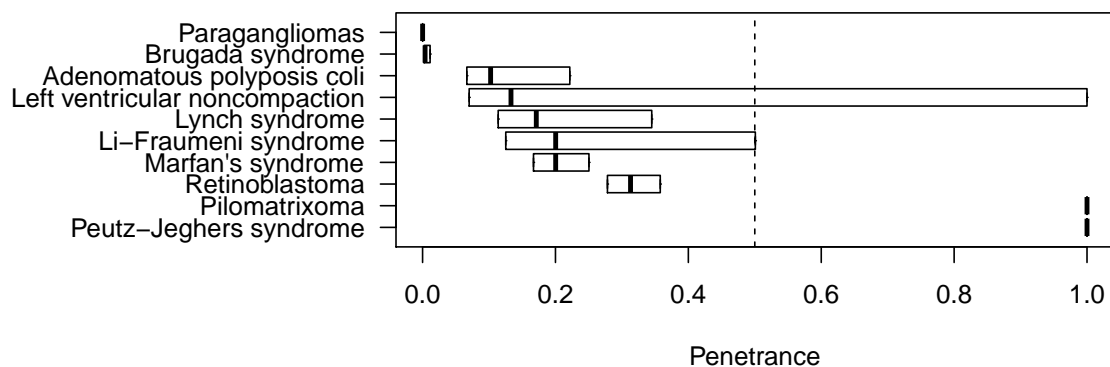
Note that this is only possible in the dozen or so cases where a prevalence range was given, rather than a point estimate.

Disease	Prevalence_Ratio
Retinoblastoma	1.3
Marfan's syndrome	1.5
Lynch syndrome	3
Adenomatous polyposis coli	3.3
Li-Fraumeni syndrome	4
Parangangliomas	4
Pilomatrixoma	4
Brugada syndrome	5
Peutz-Jeghers syndrome	12
Left ventricular noncompaction	18.6

Penetrance Estimates for Prevalence Ranges, $P(V|D) = 0.1$



Penetrance Estimates for Prevalence Ranges, $P(V|D) = 1$



Fraction with max theoretical penetrance $< P$

