ACMG-ClinVar Markdown File

James Diao

October 25, 2016

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

Steps

1. Download, Transform, and Load Data
2. Plot Summary Statistics Across Populations
3. Compute Penetrance Estimates

Download, Transform, and Load Data

1. Scrape ACMG gene panel from ClinVar

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
```

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 and columns):
```

Table continues below

VAR_ID	CHROM	POS	ID	REF	ALT	CLNSIG	CLNDBN	CLNDSDBID	INTERP
1_955597_G_T	1	955597	rs115173026	G	T	2	not_specified	CN169374	FALSE
1_955619_G_C	1	955619	rs201073369	G	C	255	not_specified	CN169374	FALSE
1_957605_G_A	1	957605	rs756623659	G	A	5	Congenital_myasthenic_syndrome	C0751882:ORPHA590	TRUE

3. Download 1000 Genomes VCFs and collect ACMG-56 regions (via tabix)

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 Temp_Files
```

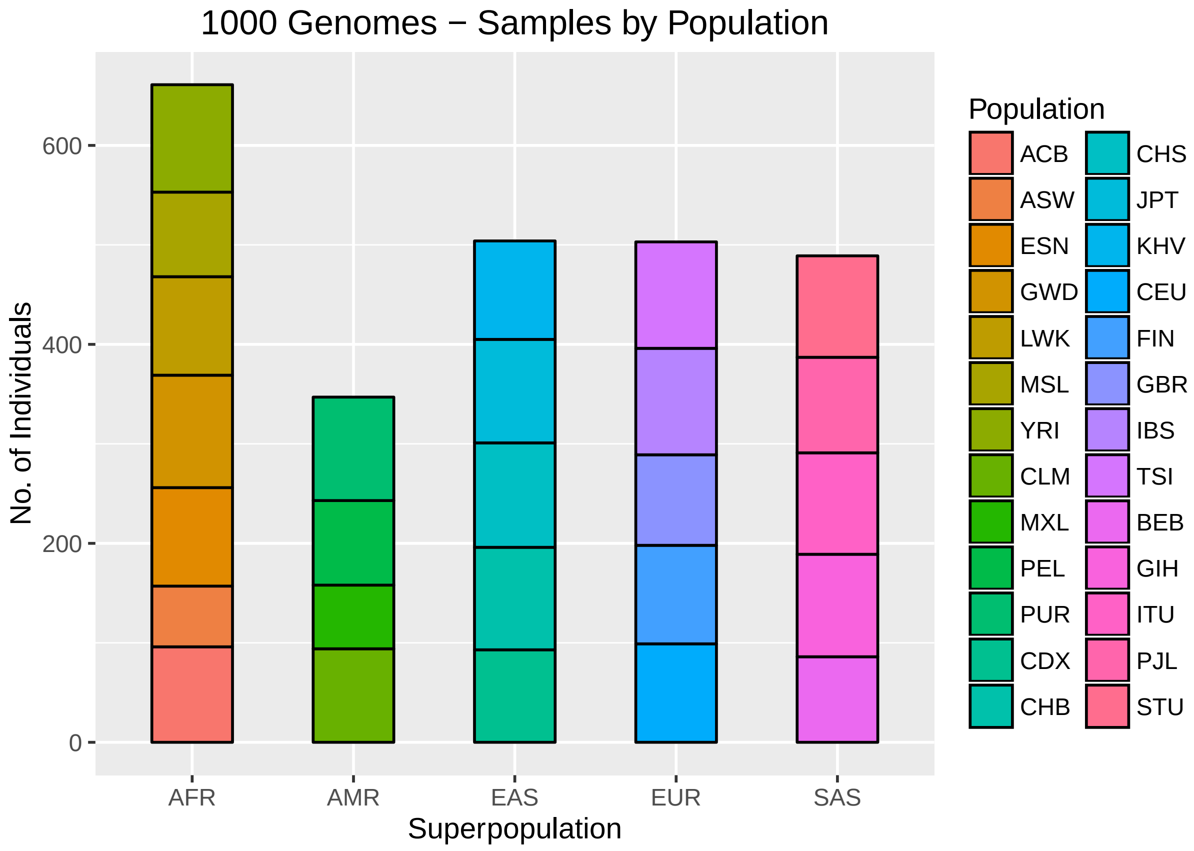
4. Access 1000 Genomes Phase 3 Populations Map

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
NA19144	YRI	AFR	male
NA19726	MXL	AMR	male
NA19762	MXL	AMR	male
HG00631	CHS	EAS	male
HG00766	CDX	EAS	female
NA20510	TSI	EUR	male
HG03636	PJL	SAS	male
HG02786	PJL	SAS	male
NA21088	GIH	SAS	female
HG03937	BEB	SAS	female

```
## Population Distribution
```



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: 139335 x 2516 (selected rows and columns):
```

GENE	AF_1000G	VAR_ID	CHROM	POS	ID	REF	ALT	HG00096	HG00097	HG00099	HG00100	HG00101	HG00102
APC	0.0001997	5_112043211_A_G	5	1.12e+08	rs554351451	A	G	0	0	0	0	0	0
APC	0.0001997	5_112043231_G_A	5	1.12e+08	rs575784409	G	A	0	0	0	0	0	0
APC	0.005391	5_112043234_C_T	5	1.12e+08	rs115658307	C	T	0	0	0	0	0	0
APC	0.0001997	5_112043252_G_A	5	1.12e+08	rs558562104	G	A	0	0	0	0	0	0
APC	0.008786	5_112043263_C_T	5	1.12e+08	rs138386816	C	T	0	0	0	0	0	0

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: 58873 x 45 (selected rows and columns):
```

GENE	AF_EXAC	AF_EXAC_AFR	AF_EXAC_AMR	AF_EXAC_EAS	AF_EXAC_EUR	AF_EXAC_SAS	VAR_ID	CHROM	POS	ID	REF	ALT
APC	8.13e-05	0	0	0	0	0.0001313	5_112043365_G_C	5	1.12e+08	.	G	C
APC	8.131e-05	0	0	0	0	0.0001313	5_112043382_A_G	5	1.12e+08	.	A	G
APC	0.1112	0.07979	0.1022	0	0.1063	0.1185	5_112043384_T_G	5	1.12e+08	rs78429131	T	G
APC	8.131e-05	0	0	0	0	0.0001313	5_112043392_C_T	5	1.12e+08	.	C	T
APC	8.134e-05	0	0	0	0	0.0001313	5_112043412_C_G	5	1.12e+08	.	C	G

7. Merge ClinVar with 1000 Genomes and ExAC (keep pathogenic variants)

```
## 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

8. Compare 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 /Temp_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)
```

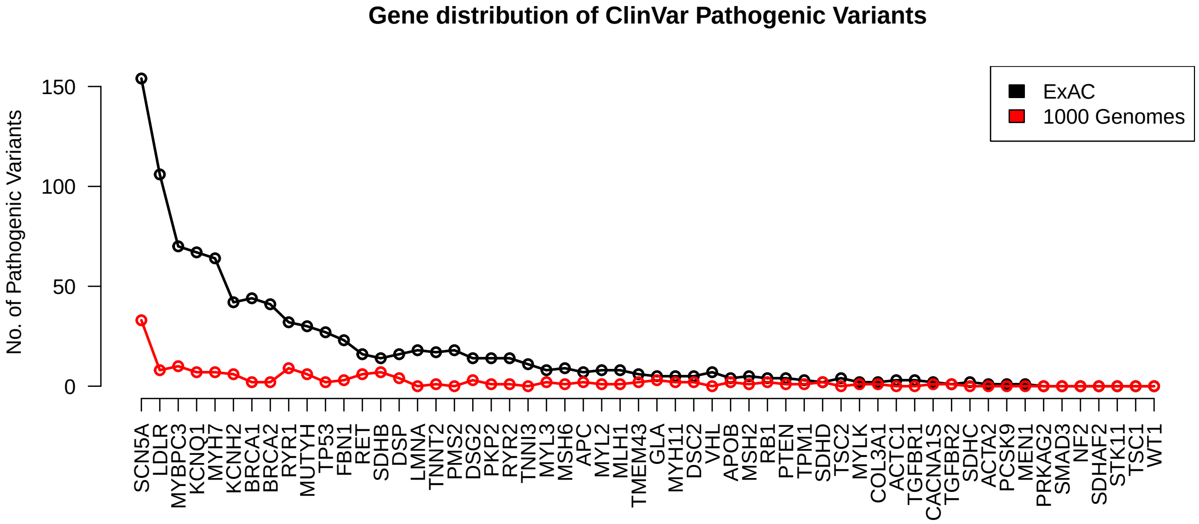
VAR_ID	Gene(s)	Condition(s)	Frequency
X_100652891_C_G	GLA	Fabry disease	GMAF:0.00050(G)
11_47374186_C_G	MYBPC3	Primary familial hypertrophic cardiomyopathy	GMAF:0.00020(G)
11_47355233_C_G	MYBPC3	Familial hypertrophic cardiomyopathy 4	GMAF:0.00020(G)
11_47364162_C_G	MYBPC3	Familial hypertrophic cardiomyopathy 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
In LP/P-ClinVar VCF & ACMG & ExAC	48
In LP/P-ClinVar VCF & ACMG & 1000 Genomes	9
In LP/P-ClinVar VCF & ACMG & ExAC & 1000 Genomes	8

Plot Summary Statistics Across Populations

1. Gene distribution of ClinVar Pathogenic Variants



2. Overall Non-Reference Sites

For 1000 Genomes

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

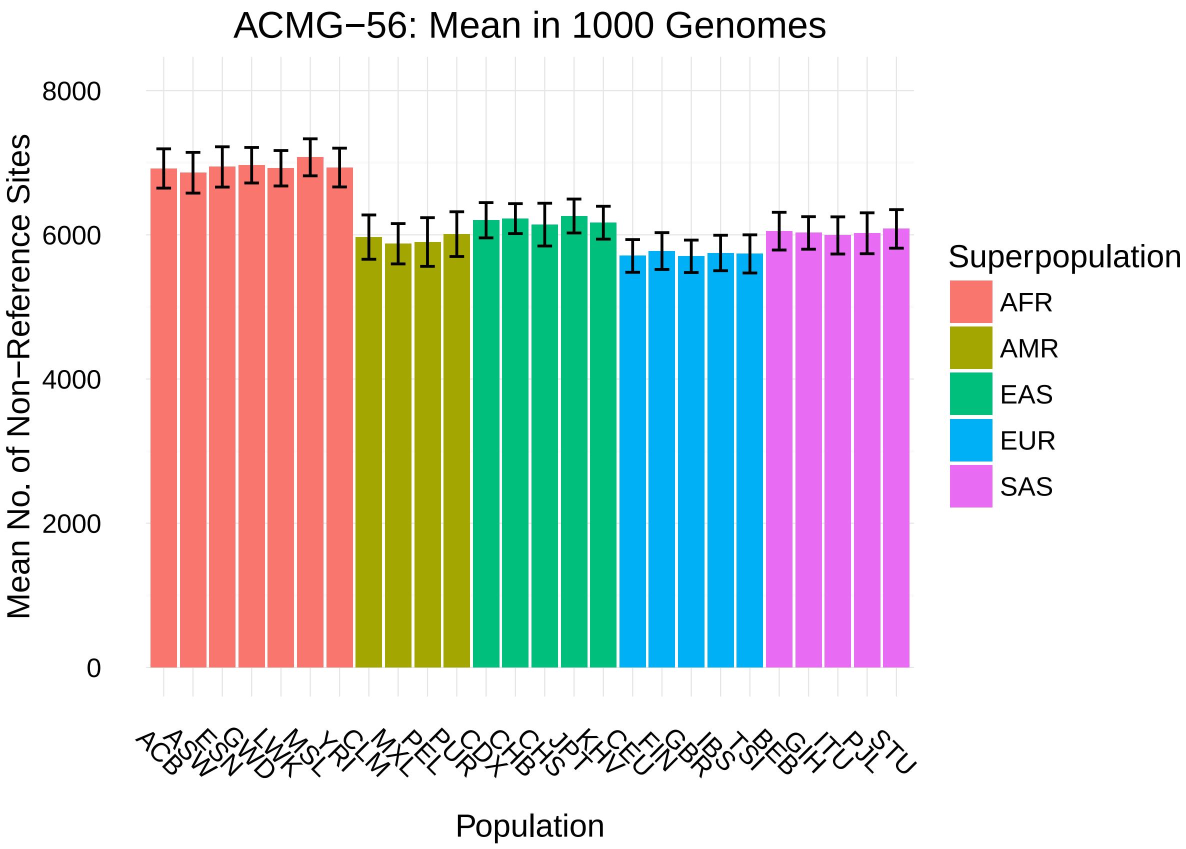
Ex: the genotype of 3 variants in 3 people looks like this:

```
##           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.

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.

Note: this is not true in some rare cases, e.g., when multiple variants share the same position.

Since we have the allele frequencies for each superpopulation, we can estimate $E(V)$ for each superpopulation.

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

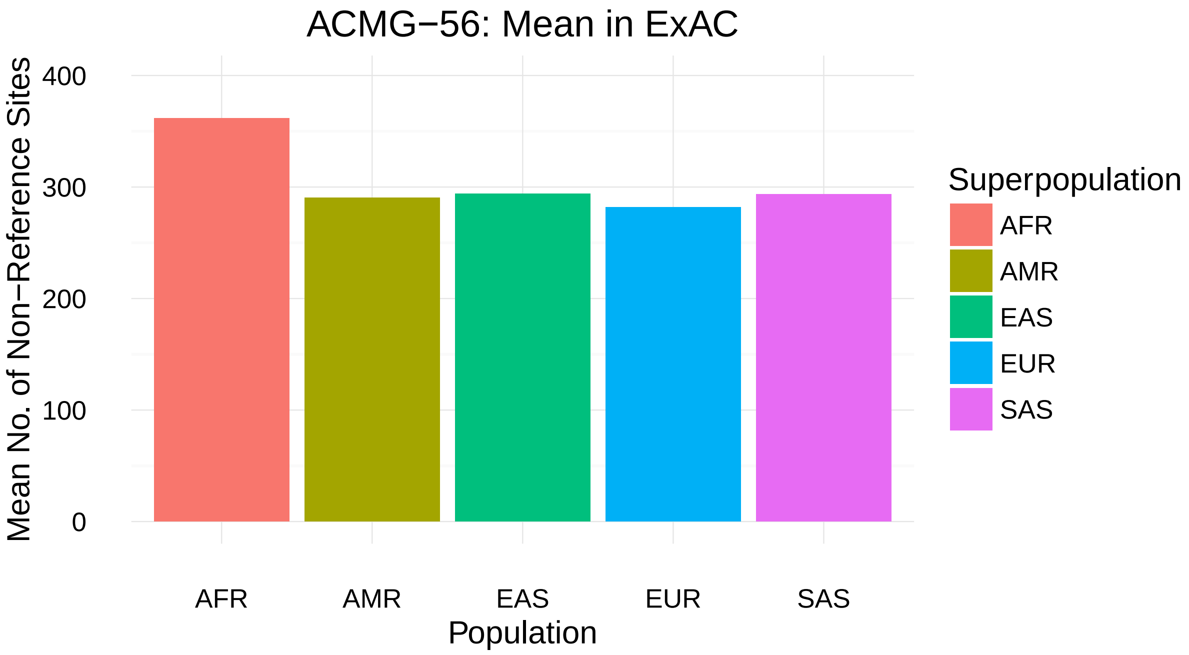
```
##              AF_EXAC_AFR AF_EXAC_AMR AF_EXAC_EAS AF_EXAC_EUR AF_EXAC_SAS
## Variant 1            0.1         0.2         0.0         0.0         0.3
## Variant 2            0.2         0.0         0.3         0.0         0.1
## Variant 3            0.0         0.0         0.1         0.1         0.2
```

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:

```
##              AF_EXAC_AFR AF_EXAC_AMR AF_EXAC_EAS AF_EXAC_EUR AF_EXAC_SAS
## Variant 1           0.19        0.36        0.00        0.00        0.51
## Variant 2           0.36        0.00        0.51        0.00        0.19
## Variant 3           0.00        0.00        0.19        0.19        0.36
```

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

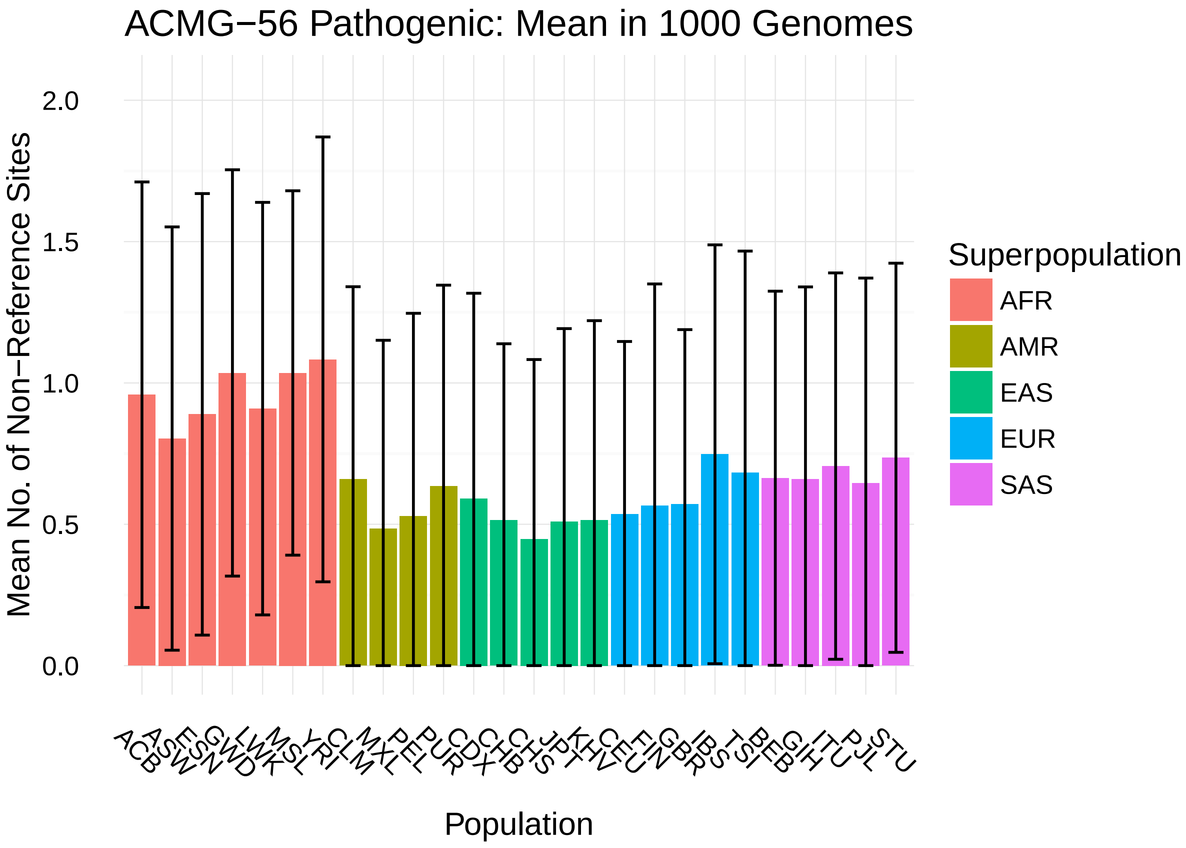
```
## AF_EXAC_AFR AF_EXAC_AMR AF_EXAC_EAS AF_EXAC_EUR AF_EXAC_SAS
##        0.55        0.36        0.70        0.19        1.06
```

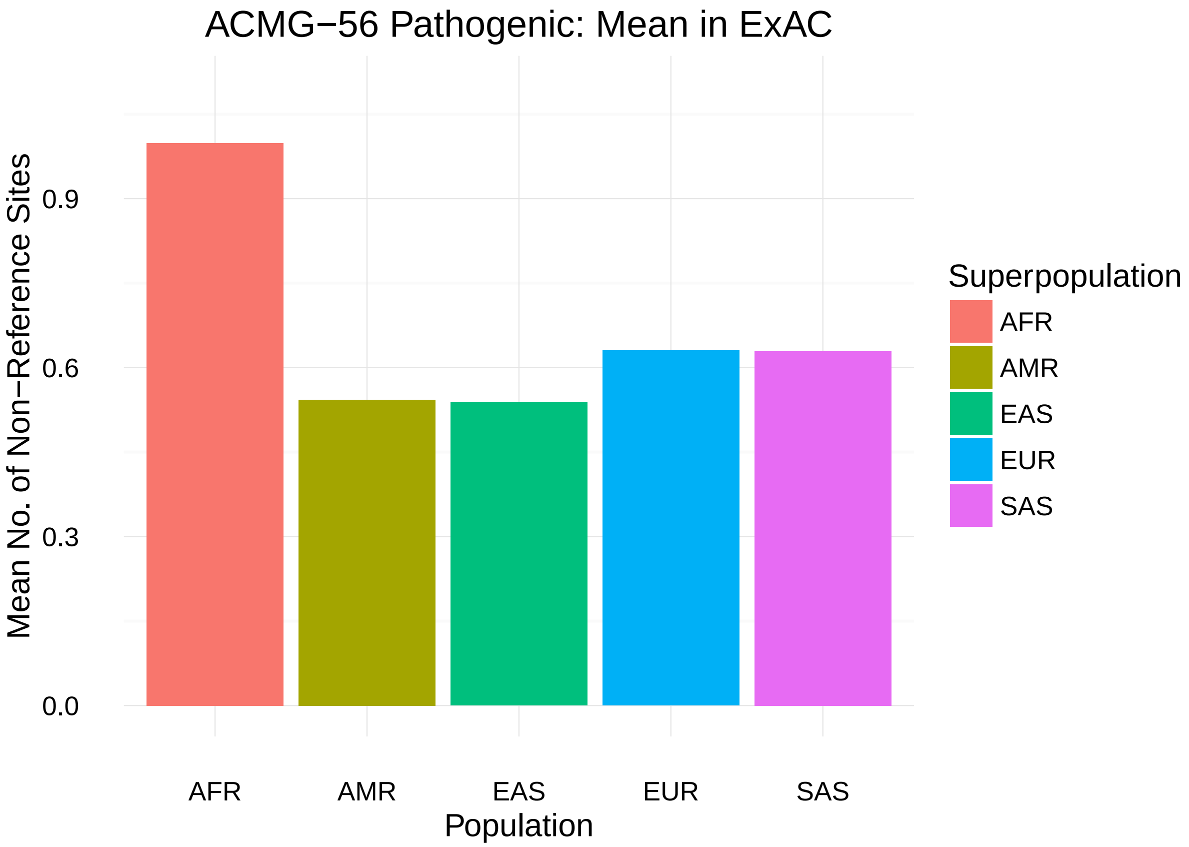


3. Pathogenic Non-Reference Sites

For 1000 Genomes and ExAC

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





4. Fraction of 1000 Genomes Individuals with Pathogenic Sites

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.

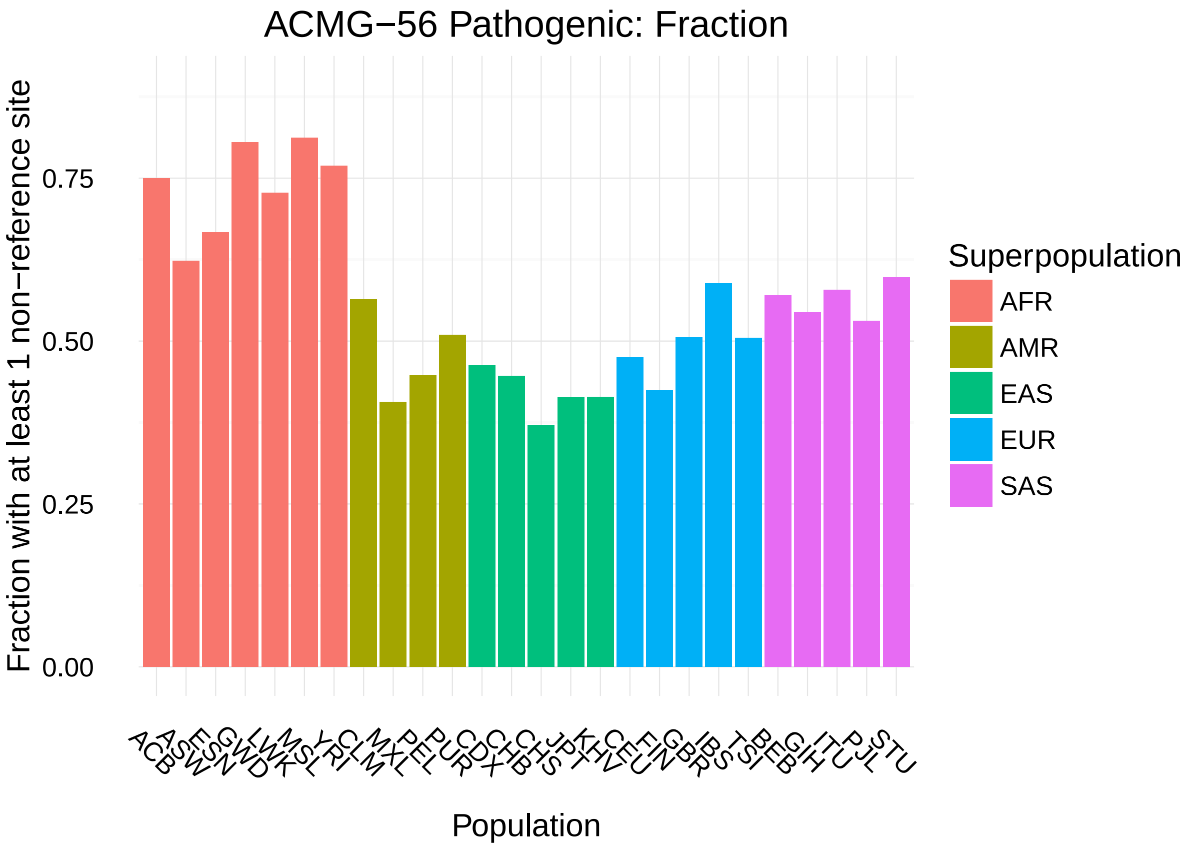
Ex: the genotype of 3 variants in 3 people looks like this:

```
##           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
```



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:

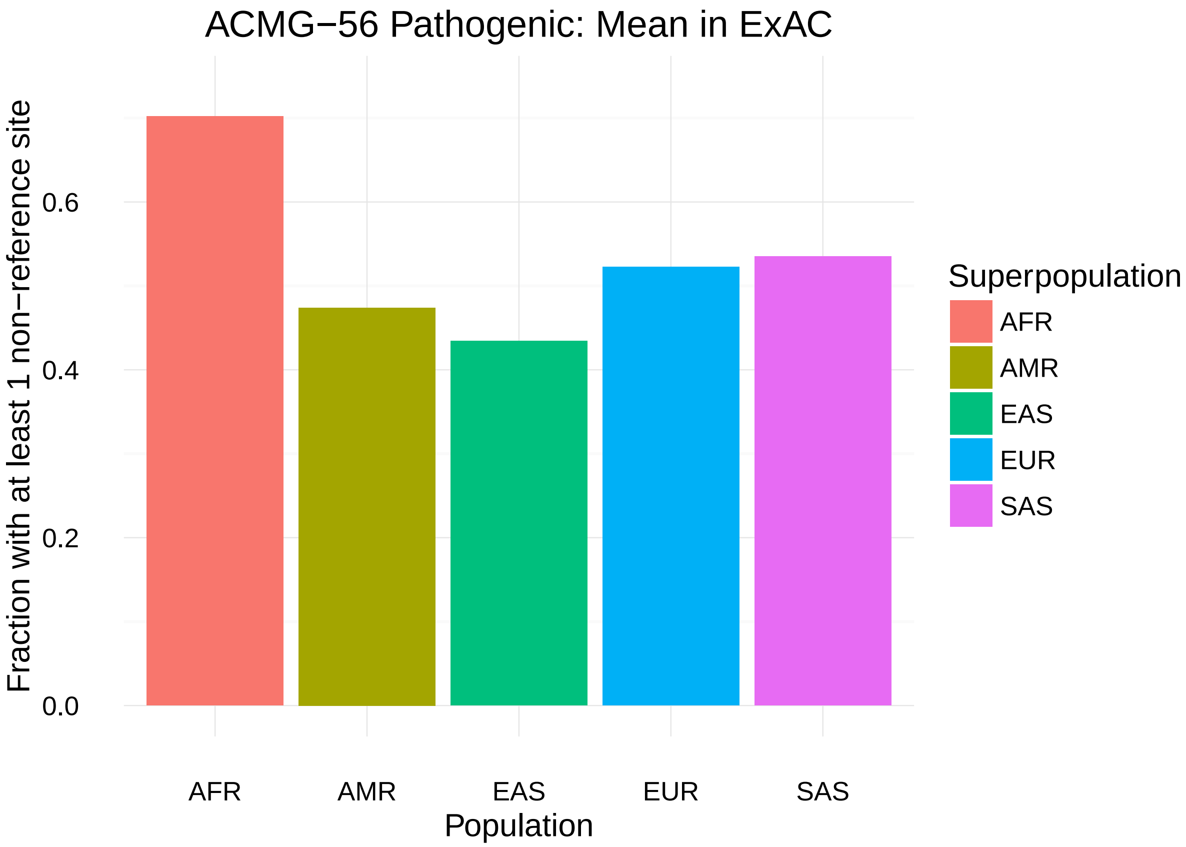
```
##           AF_EXAC_AFR AF_EXAC_AMR AF_EXAC_EAS AF_EXAC_EUR AF_EXAC_SAS
## Variant 1         0.1         0.2         0.0         0.0         0.3
## Variant 2         0.2         0.0         0.3         0.0         0.1
## Variant 3         0.0         0.0         0.1         0.1         0.2
```

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:

```
##           AF_EXAC_AFR AF_EXAC_AMR AF_EXAC_EAS AF_EXAC_EUR AF_EXAC_SAS
## Variant 1        0.19        0.36        0.00        0.00        0.51
## Variant 2        0.36        0.00        0.51        0.00        0.19
## Variant 3        0.00        0.00        0.19        0.19        0.36
```

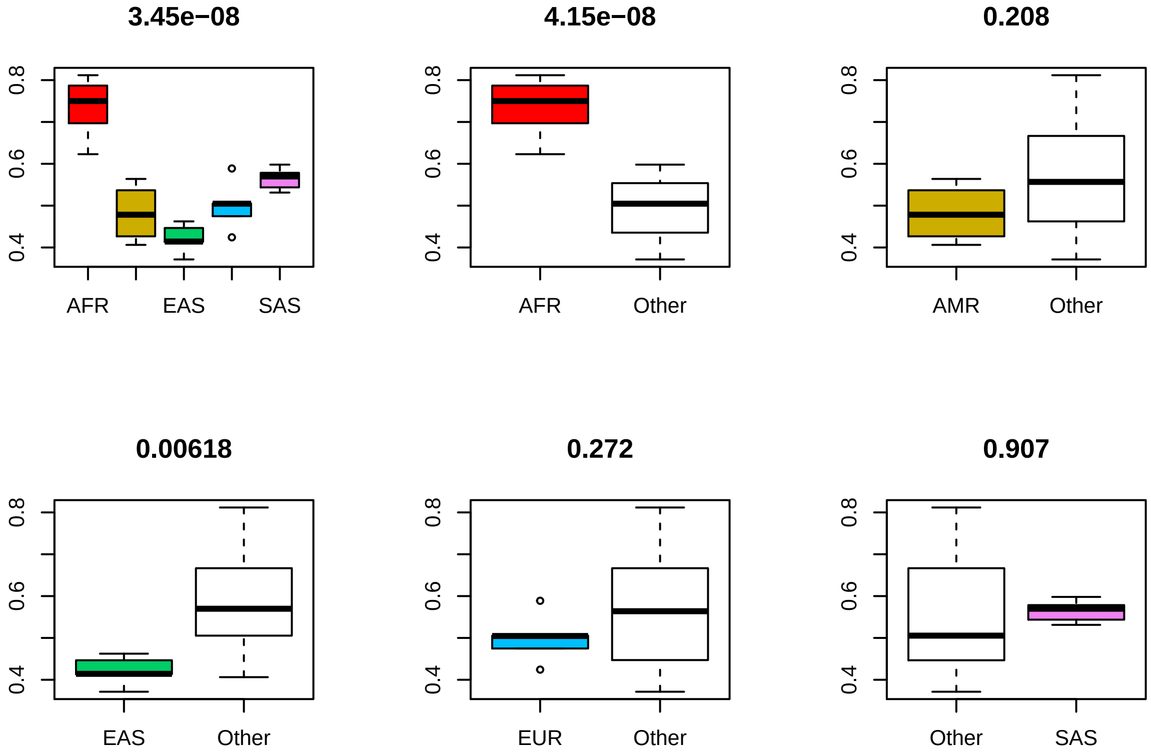
The expected (mean) number of non-reference sites is given by $1 - \prod(1 - AF)^2$.

```
## AF_EXAC_AFR AF_EXAC_AMR AF_EXAC_EAS AF_EXAC_EUR AF_EXAC_SAS
##    0.481600    0.360000    0.603100    0.190000    0.745984
```

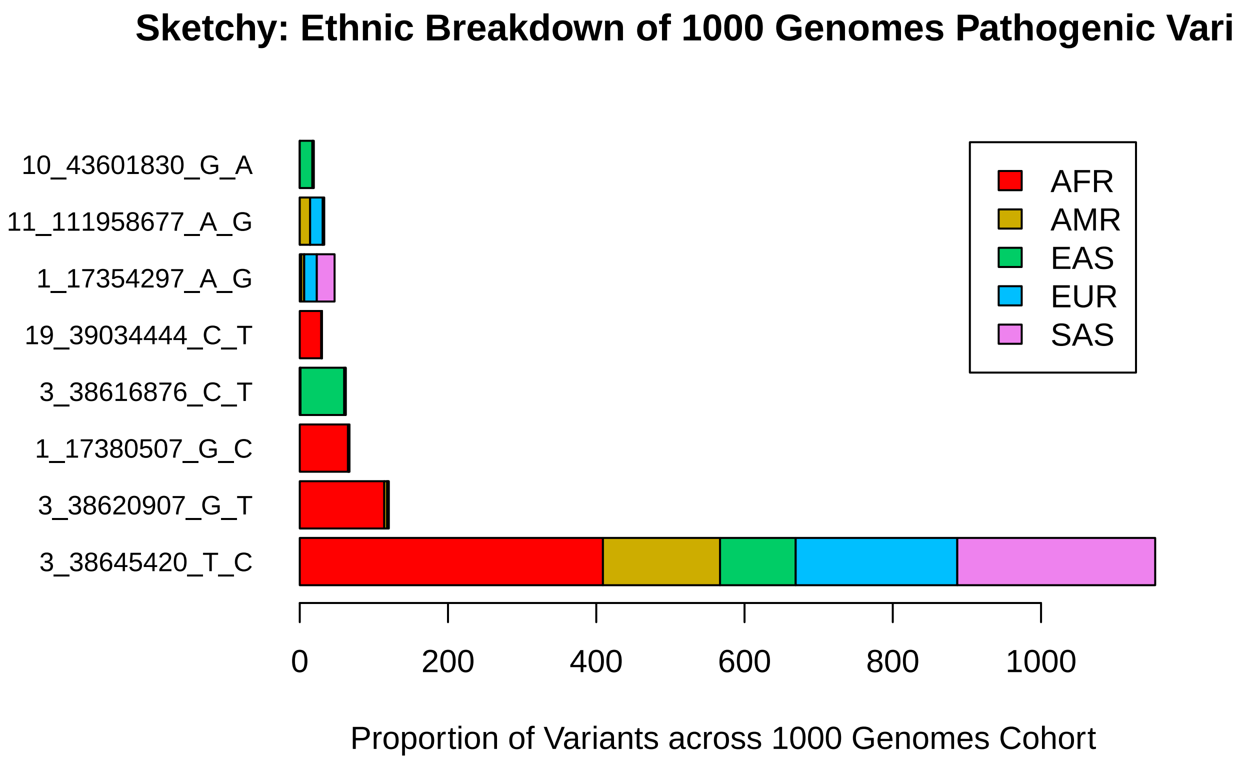

ACMG−56 Pathogenic: Mean in ExAC
Fraction with at least 1 non−reference site
0.0
0.2
0.4
0.6
AFR
AMR
EAS
EUR
SAS
Population
Superpopulation
AFR
AMR
EAS
EUR
SAS

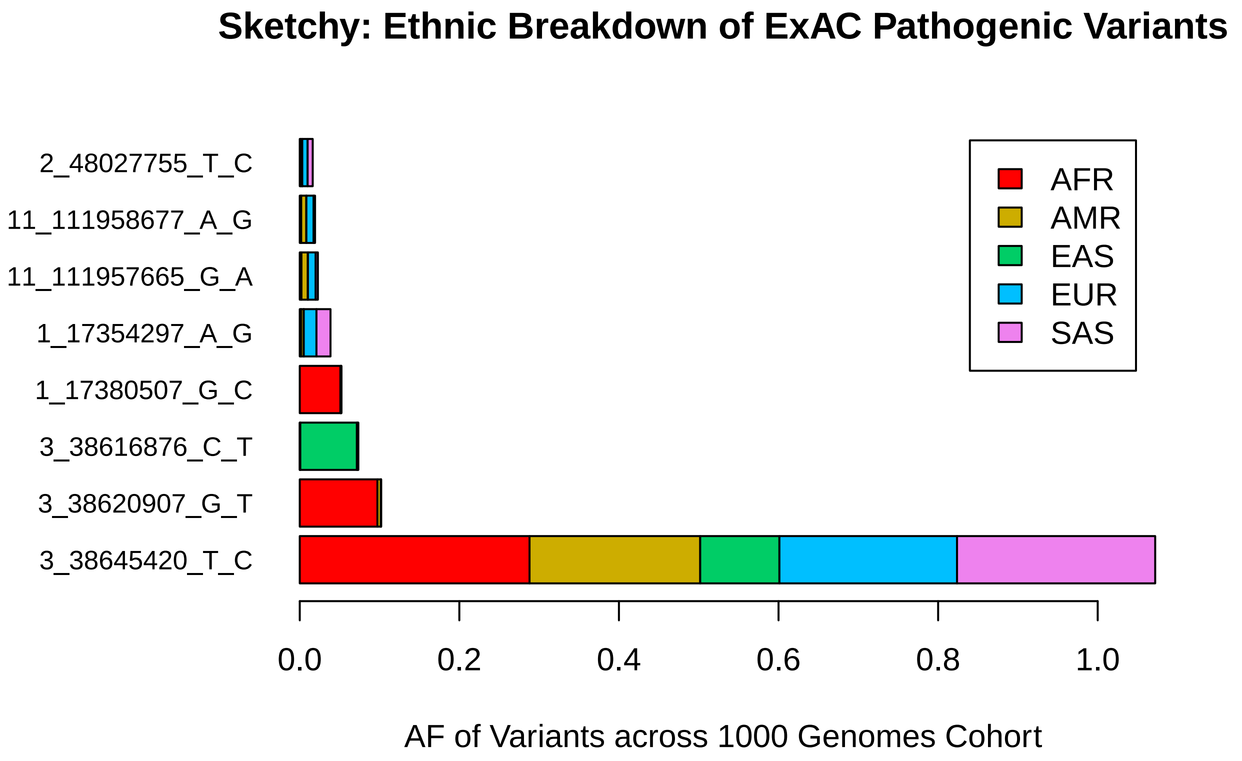
5. Test Statistics

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.



6. Ethnic Breakdown of 1000 Genomes Pathogenic Variants


Sketchy: Ethnic Breakdown of 1000 Genomes Pathogenic Vari
10_43601830_G_A
11_111958677_A_G
1_17354297_A_G
19_39034444_C_T
3_38616876_C_T
1_17380507_G_C
3_38620907_G_T
3_38645420_T_C
0
200
400
600
800
1000
Proportion of Variants across 1000 Genomes Cohort
🟥 AFR
🟨 AMR
🟩 EAS
🟦 EUR
🟪 SAS


Sketchy: Ethnic Breakdown of ExAC Pathogenic Variants
2_48027755_T_C
11_111958677_A_G
11_111957665_G_A
1_17354297_A_G
1_17380507_G_C
3_38616876_C_T
3_38620907_G_T
3_38645420_T_C
0.0
0.2
0.4
0.6
0.8
1.0
AF of Variants across 1000 Genomes Cohort
🟥 AFR
🟨 AMR
🟩 EAS
🟦 EUR
🟪 SAS

Penetrance Estimates

1. Import Literature Search Disease Prevalence Data

```
##    Gene          Disease Inverse.P1 Inverse.P2 Inverse.Prevalence Region
## 5 SCN5A Brugada syndrome      10000       2000               4472  World
##                                          URL
## 5 http://www.ncbi.nlm.nih.gov/pubmed/17038146
##                                journal year sample.size first.author
## 5 Pacing and Clinical Electrophysiology 2006        <NA> Antzelevitch
##   subset citations date.accessed                         issue
## 5   <NA>        11    19-07-2016 refers to other (uncited) studies
```

2. Manually curated disease keywords from CLNDBN

```
##  [1] adenomatous                     aneurysm
##  [3] arrhythmogenic;dreifuss         breast;ovarian
##  [5] brugada;gardner                 tachycardia
##  [7] dilated                         ehler
##  [9] fabry                           hypercholesterolemia
## [11] hypertrophic                    medullary
## [13] noncompaction                   Fraumeni
## [15] Loeys;Dietz                     QT
## [17] lynch;endometrial               hyperthermia
## [19] Marfan                          neoplasia;men2a
## [21] MYH;colon                       neurofibromatosis
## [23] paraganglioma;pheochromocytoma peutz;jeghers
## [25] pilomatrixoma                   Cowden;PTEN;hamartoma;Merkel
## [27] retinoblastoma                  tuberous
## [29] Hippel;Lindau                   Wilms
```

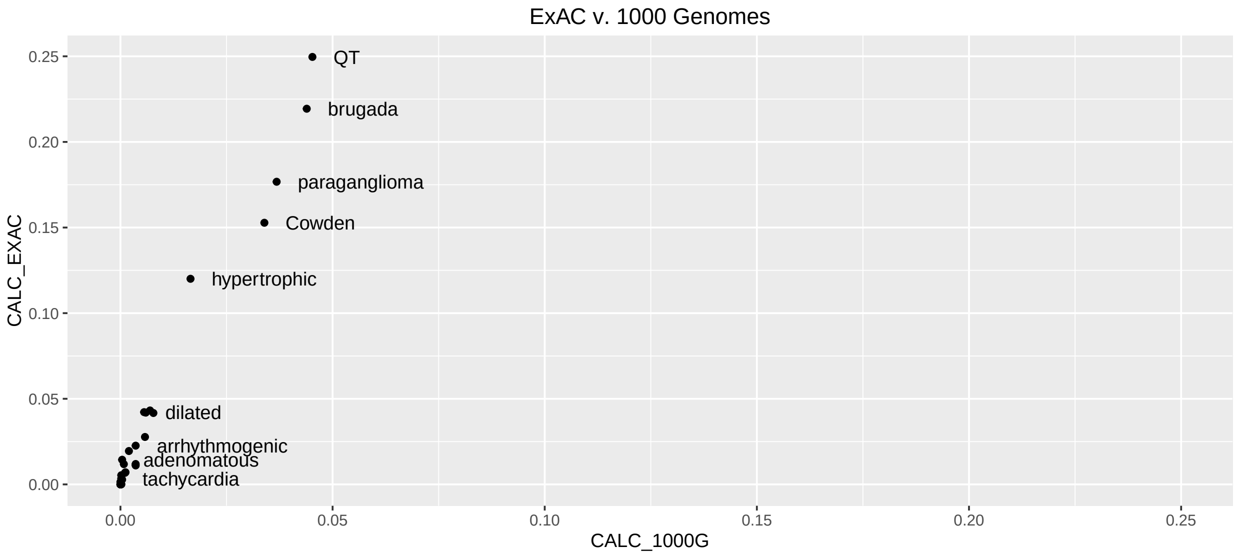
3. Aggregate allele frequencies of all variants associated with each disease - AF(disease)

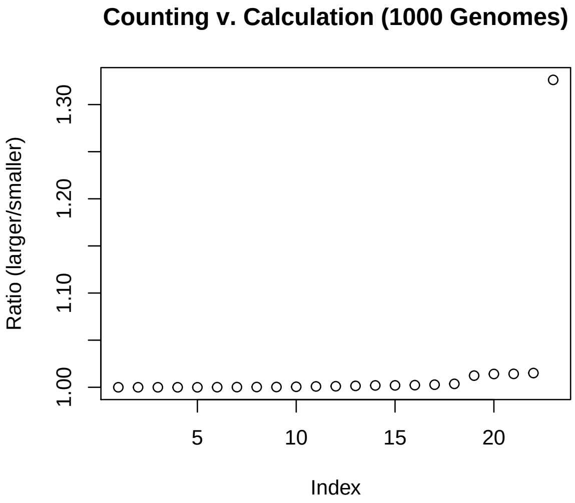
We define AF(disease) as the probability of having at least 1 variant associated with the disease. This can be computed in two ways: (a) By direct counting, from genotype data in 1000 Genomes. (b) AF(disease) = $1 - \prod_{variant}(1 - AF_{variant})$, from population data in ExAC. This assumes independence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

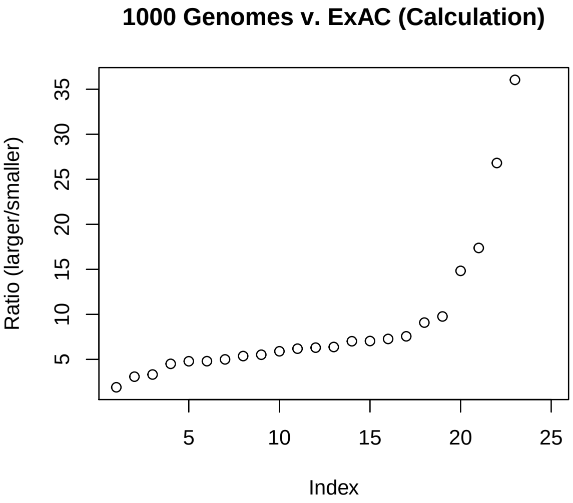
4. Comparisons of AF(disease) by dataset or method

Correlation Table:

```
##             COUNT_1000G CALC_1000G CALC_EXAC
## COUNT_1000G   1.0000000  0.9999898 0.9897085
## CALC_1000G    0.9999898  1.0000000 0.9898656
## CALC_EXAC     0.9897085  0.9898656 1.0000000
```

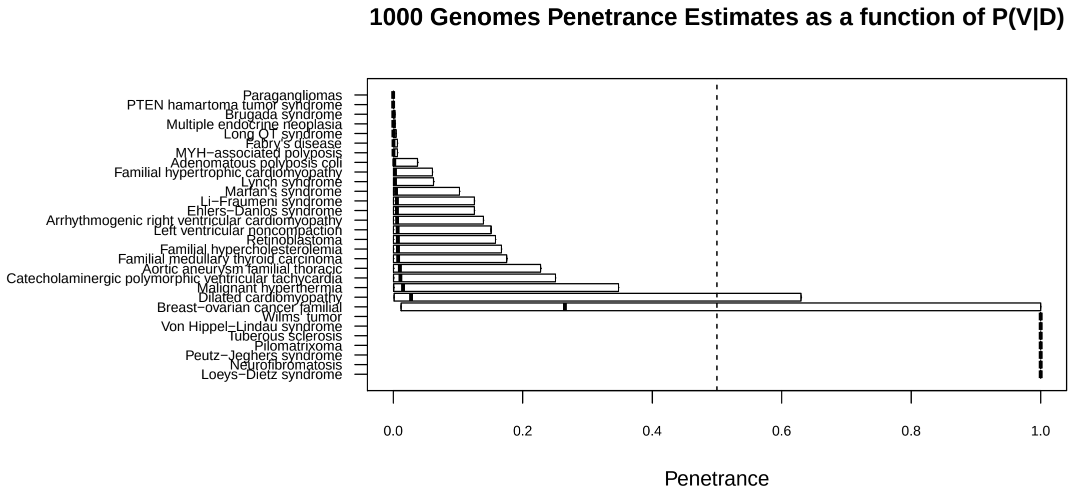




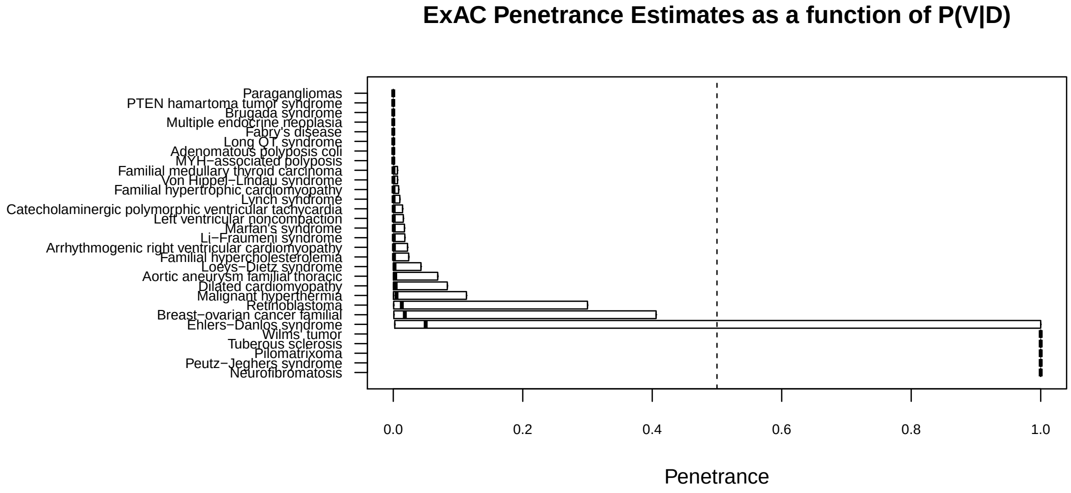


The median AF(disease) ratio between counting and calculation is: 1.001.
The median AF(disease) ratio between ExAC and 1000 Genomes is: 6.302.

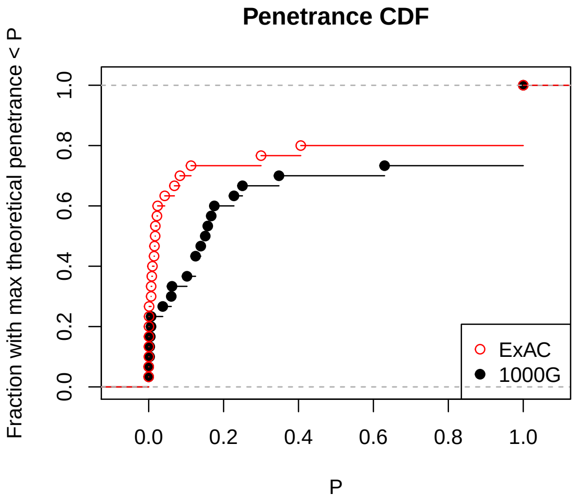
5. Penetrance as a function of allelic heterogeneity: P(V|D) = 0.001, 0.02, 0.5



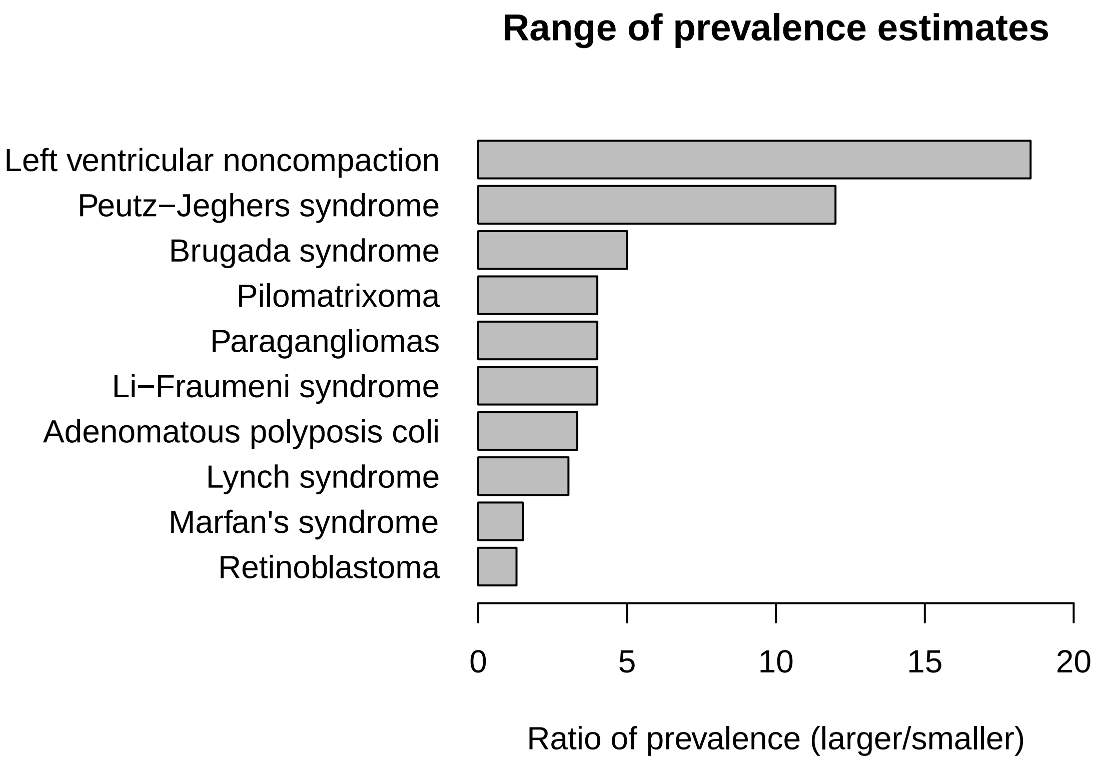
1000 Genomes Penetrance Estimates as a function of P(V|D)

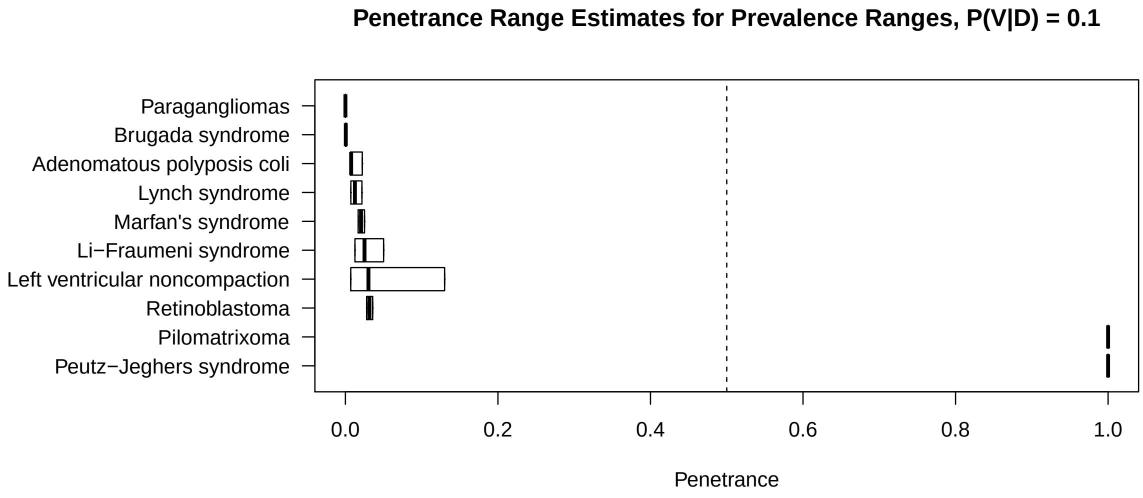


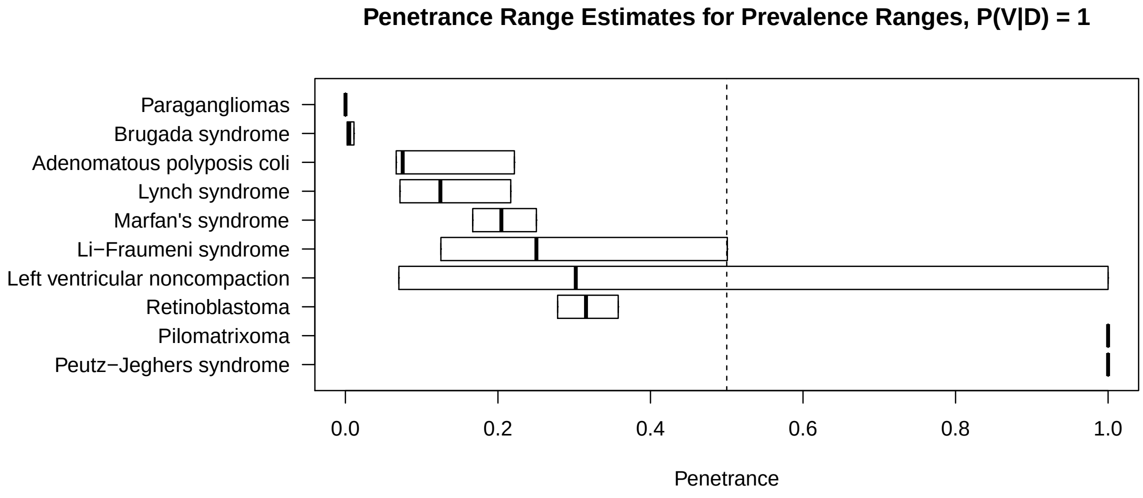
ExAC Penetrance Estimates as a function of P(V|D)

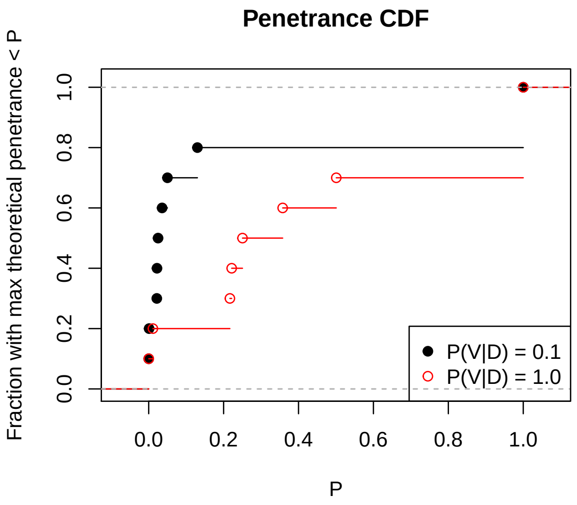


6. **Penetrance as a function of prevalence (when a range is presented)**




Penetrance Range Estimates for Prevalence Ranges, P(V|D) = 0.1
Paragangliomas
Brugada syndrome
Adenomatous polyposis coli
Lynch syndrome
Marfan's syndrome
Li−Fraumeni syndrome
Left ventricular noncompaction
Retinoblastoma
Pilomatrixoma
Peutz−Jeghers syndrome
0.0
0.2
0.4
0.6
0.8
1.0
Penetrance


Penetrance Range Estimates for Prevalence Ranges, P(V|D) = 1
Paragangliomas
Brugada syndrome
Adenomatous polyposis coli
Lynch syndrome
Marfan's syndrome
Li−Fraumeni syndrome
Left ventricular noncompaction
Retinoblastoma
Pilomatrixoma
Peutz−Jeghers syndrome
0.0
0.2
0.4
0.6
0.8
1.0
Penetrance


Penetrance CDF
Fraction with max theoretical penetrance < P
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0
P
● P(V|D) = 0.1
○ P(V|D) = 1.0

–>