# ClinVar Report

## James Diao

## November 4, 2016

## Contents

1	ollect and Merge ClinVar Data  1 Import ClinVar VCF	
2	ummary Statistics	3
	1 Fraction of Individuals with Pathogenic Non-Reference Sites	. 3
3	enetrance Estimates	4
	1 Max/Min Penetrance as a Function of $P(D)$ and $P(V D)$	. 4
	2 Penetrance Estimates by Ancestry	
	3 Empirical CDFs for All Penetrance Plots	. 7
	4 Comparing Mean Penetrance between ExAC and 1000 Genomes	. 7
	cing ClinVar input from: clinvar_2013-01-14.vcf ling output to: Report_2013-01-14.pdf	

## 1 Collect and Merge ClinVar Data

## 1.1 Import ClinVar VCF

## Processed ClinVar data frame  $67257 \times 14$  (selected rows/columns):

## 1.2 Merge ClinVar with 1000 Genomes and ExAC

#### ## Breakdown of ClinVar Variants

Subset_ClinVar	Number_of_Variants
Total ClinVar	67257
LP/P-ClinVar	12607
LP/P-ClinVar & ACMG	915
LP/P-ClinVar & ACMG & ExAC	206
LP/P-ClinVar & ACMG & 1000	65
Genomes	

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

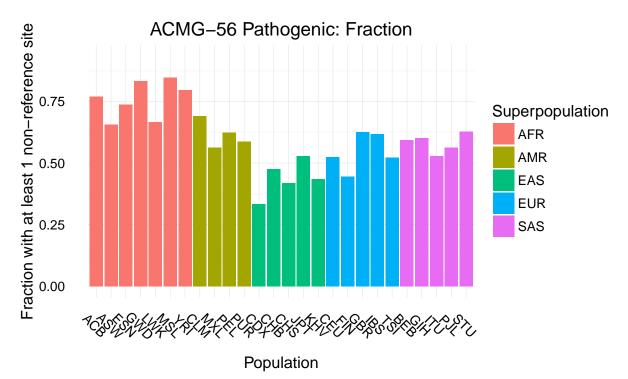
Subset_1000_Genomes	Number_of_Variants
Total 1000_Genomes & ACMG	139335
1000_Genomes & ACMG & ClinVar 1000 Genomes & ACMG &	1214 65
LP/P-ClinVar	05

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

Subset_ExAC	Number_of_Variants
Total ExAC & ACMG	58873
ExAC & ACMG & ClinVar	2256
ExAC & ACMG & LP/P-ClinVar	206

## 2 Summary Statistics

## 2.1 Fraction of Individuals with Pathogenic Non-Reference Sites



ACMG-56 Pathogenic: Mean in ExAC

Superpopulation

AFR

AMR

EAS

Population

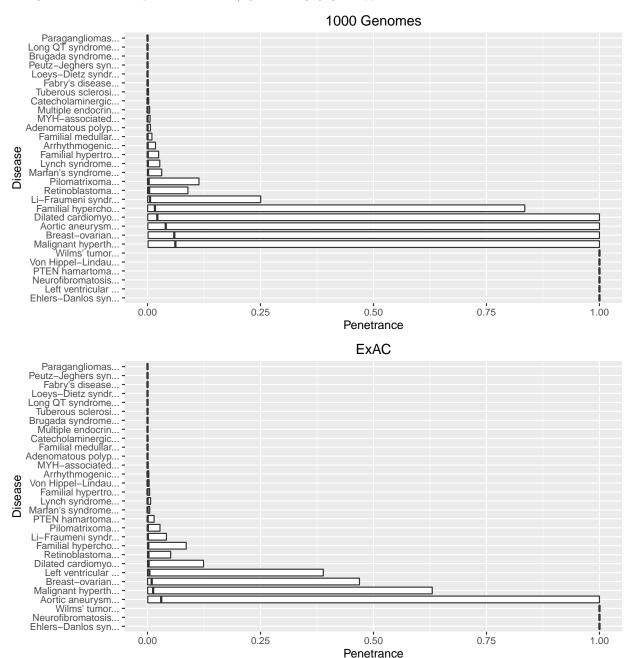
AFR

SAS

#### 3 Penetrance Estimates

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

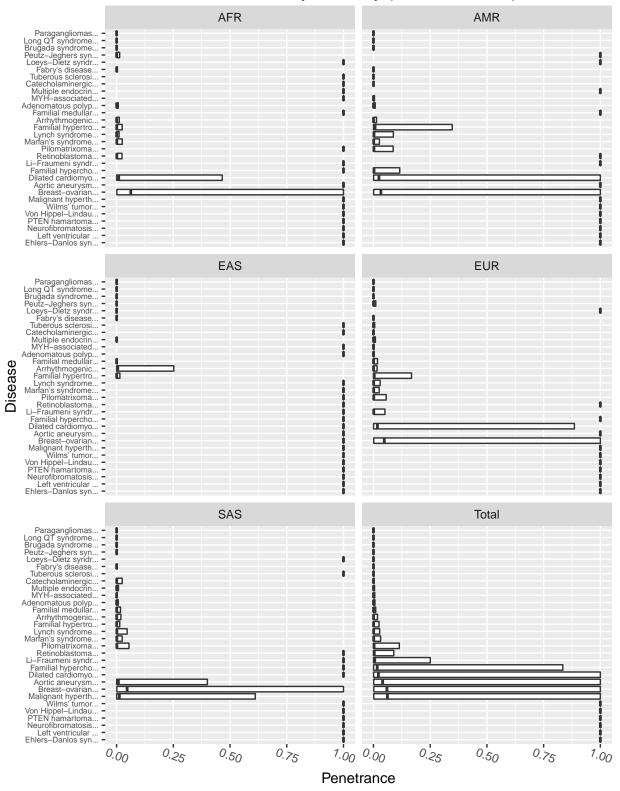
The left end of the boxplot indicates P(D) AND P(V|D) = lower value, the bold line in the middle indicates P(D) AND  $P(V|D) = geometric\_mean(values)$ , the right end of the boxplot indicates P(D) AND P(V|D) = upper value.



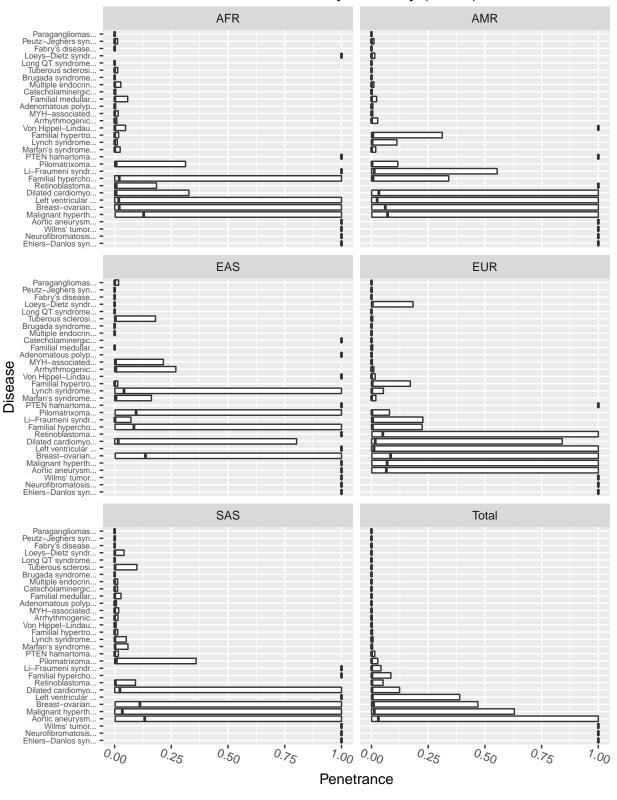
Note: Prevalence ranges of 5x were assumed for all point estimates of prevalence. For example: a point estimate of 0.022 would be given the range 0.01-0.05.

#### 3.2 Penetrance Estimates by Ancestry

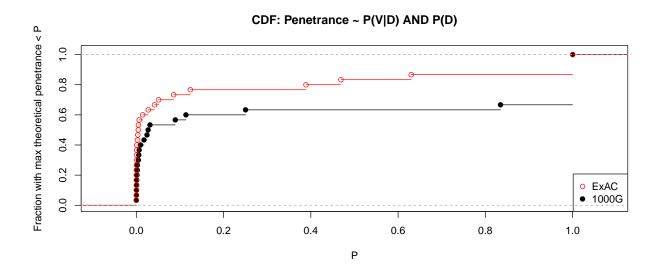
## Penetrance by Ancestry (1000 Genomes)



### Penetrance by Ancestry (ExAC)



### 3.3 Empirical CDFs for All Penetrance Plots



#### 3.4 Comparing Mean Penetrance between ExAC and 1000 Genomes

#### Penetrance Means: ExAC v. 1000 Genomes Aortic aneurysm... Malignant hyperth... 1e-02 -Breast-ovarian... Dilated cardiomyo... Familial hypercho... 1e-03 -Penetrance\_ExAC Li-Fraumeni syndr... Pilomatrixoma Lynch syndrome.. Marfan's syndrome... Familial hypertro... MYH-associated. Arrhythmogenic... Adenomatous polyp... Familial medullar... 1e-05 - Brugada syndrome Loeys-Dietz syndr... Multiple endocrin... Long QT syndrome. Tuberous sclerosi... Peutz-Jeghers syn... Fabry's disease... 1e-06 **-**Paragangliomas... 1e-05 1e-04 1e-06 1e-03 1e-02 1e-01 1e+00

The Pearson correlation is 0.74. Max penetrance values computed using 1000 Genomes are 2-fold larger than those computed using ExAC.

Penetrance\_1000\_Genomes