ClinVar Report

James Diao

November 4, 2016

Contents

1	1.1	lect and Merge ClinVar Data Import ClinVar VCF		
2	Sum 2.1	nmary Statistics Fraction of Individuals with Pathogenic Non-Reference Sites	3	
3		netrance Estimates	4	
	3.1	Max/Min Penetrance as a Function of $P(D)$ and $P(V D)$	4	
	3.2	Penetrance Estimates by Ancestry	5	
	3.3	Empirical CDFs for All Penetrance Plots	7	
	3.4	Comparing Mean Penetrance between ExAC and 1000 Genomes	7	
		ng ClinVar input from: clinvar_2016-04-05.vcf ng output to: Report_2016-04-05.pdf		

1 Collect and Merge ClinVar Data

1.1 Import ClinVar VCF

Processed ClinVar data frame 100642 x 14 (selected rows/columns):

1.2 Merge ClinVar with 1000 Genomes and ExAC

Breakdown of ClinVar Variants

Subset_ClinVar	Number_of_Variants
Total ClinVar	100642
LP/P-ClinVar	31042
LP/P-ClinVar & ACMG	6289
LP/P-ClinVar & ACMG & ExAC	955
LP/P-ClinVar & ACMG & 1000	168
Genomes	

Breakdown of ACMG-1000 Genomes Variants

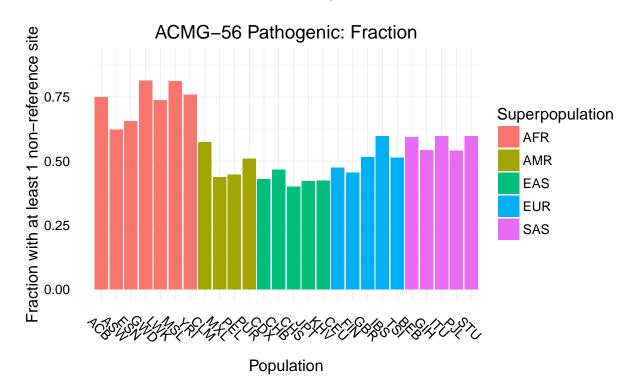
Subset_1000_Genomes	Number_of_Variants
Total 1000_Genomes & ACMG	139335
$1000_Genomes \ \& \ ACMG \ \& \ ClinVar$	4050
1000_Genomes & ACMG &	168
LP/P-ClinVar	

Breakdown of ACMG-ExAC Variants

Subset_ExAC	Number_of_Variants
Total ExAC & ACMG	58873
ExAC & ACMG & ClinVar	8091
ExAC & ACMG & LP/P-ClinVar	955

2 Summary Statistics

2.1 Fraction of Individuals with Pathogenic Non-Reference Sites



ACMG-56 Pathogenic: Mean in ExAC

Superpopulation

AFR

AMR

EAS

Population

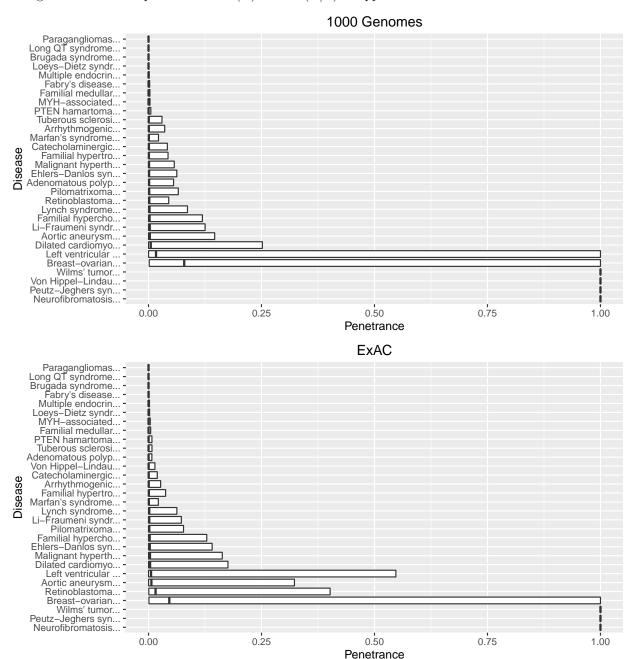
AFR

SAS

3 Penetrance Estimates

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

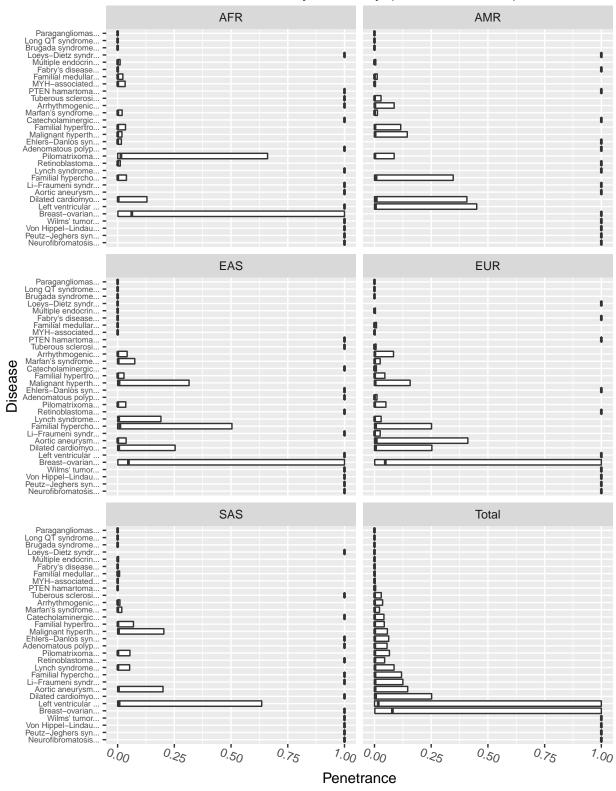
The left end of the boxplot indicates P(D) AND P(V|D) = lower value, the bold line in the middle indicates P(D) AND $P(V|D) = geometric_mean(values)$, the right end of the boxplot indicates P(D) AND P(V|D) = upper value.



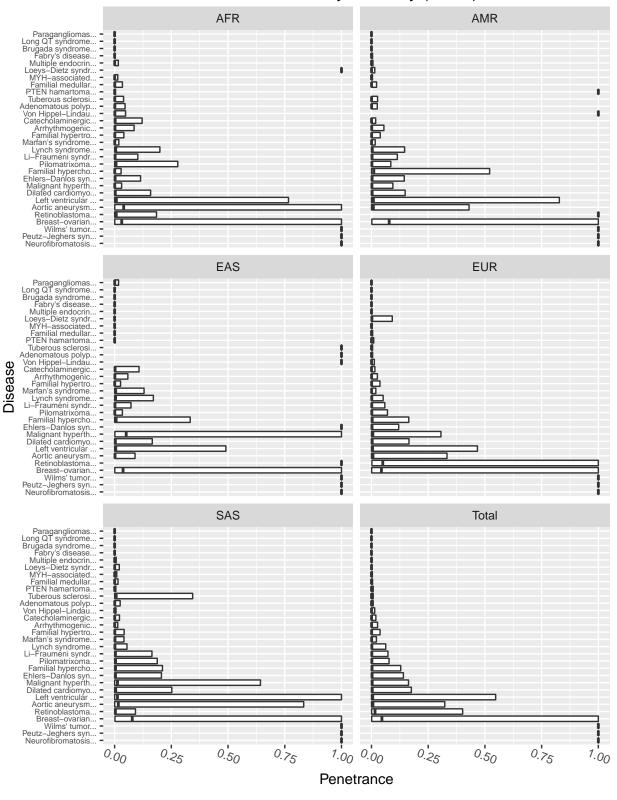
Note: Prevalence ranges of 5x were assumed for all point estimates of prevalence. For example: a point estimate of 0.022 would be given the range 0.01-0.05.

3.2 Penetrance Estimates by Ancestry

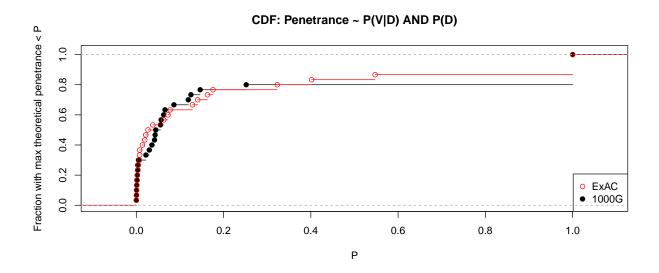




Penetrance by Ancestry (ExAC)



3.3 Empirical CDFs for All Penetrance Plots



3.4 Comparing Mean Penetrance between ExAC and 1000 Genomes

Penetrance Means: ExAC v. 1000 Genomes Breast-ovarian... Retinoblastoma... 1e-02 -Aortic aneurysm... Left ventricular ... Dilated cardiomyo... Malignant hyperth. Familial hypercho... Pilomatrixoma Penetrance_ExAC Lynch syndrome.. Marfan's syndrome... Arrhythmogenic... • Familial hypertro... Catecholaminergic... PTEN hamartoma... Adenomatous polyp... Tuberous sclerosi. MYH-associated... Multiple endocrin... Brugada syndrome.. 1e-05 -Long QT syndrome... Paragangliomas... 1e-06 -1e-02 1e+00 1e-06 1e-05 1e-03 1e-01 Penetrance_1000_Genomes

The Pearson correlation is 0.94. Max penetrance values computed using 1000 Genomes are 1.6-fold larger than those computed using ExAC.