ClinVar Report

James Diao

November 4, 2016

Contents

1		Data					
2	2 Summary Statistics						3
	2.1 Fraction of Individuals wi	th Pathogenic Non-Reference Sites	 	 	 	 	 3
3	3 Penetrance Estimates						4
	3.1 Max/Min Penetrance as a	a Function of $P(D)$ and $P(V D)$	 	 	 	 	 4
		Ancestry					
	3.3 Empirical CDFs for All P	enetrance Plots	 	 	 	 	 7
	3.4 Comparing Mean Penetra	nce between ExAC and 1000 Genomes	 	 	 	 	 7
	Sourcing ClinVar input from: Sending output to: Report_201	_					

1 Collect and Merge ClinVar Data

1.1 Import ClinVar VCF

Processed ClinVar data frame 88578 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	88578
LP/P-ClinVar	27858
LP/P-ClinVar & ACMG	5850
LP/P-ClinVar & ACMG & ExAC	980
LP/P-ClinVar & ACMG & 1000	181
Genomes	

Breakdown of ACMG-1000 Genomes Variants

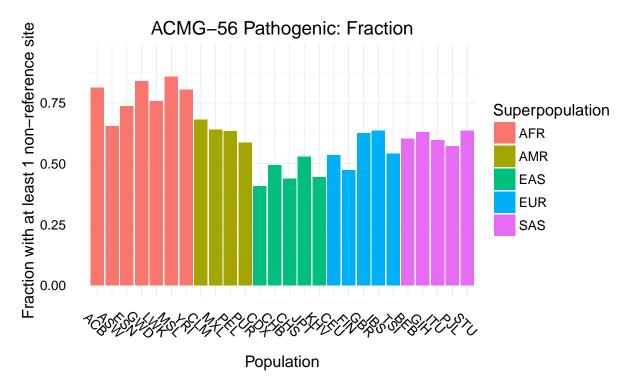
Subset_1000_Genomes	Number_of_Variants
Total 1000_Genomes & ACMG	139335
1000_Genomes & ACMG & ClinVar	3749
1000_Genomes & ACMG &	181
LP/P-ClinVar	

Breakdown of ACMG-ExAC Variants

Subset_ExAC	Number_of_Variants
Total ExAC & ACMG	58873
ExAC & ACMG & ClinVar	7504
ExAC & ACMG & LP/P-ClinVar	980

2 Summary Statistics

2.1 Fraction of Individuals with Pathogenic Non-Reference Sites



ACMG-56 Pathogenic: Mean in ExAC

Superpopulation

AFR

AMR

EAS

Population

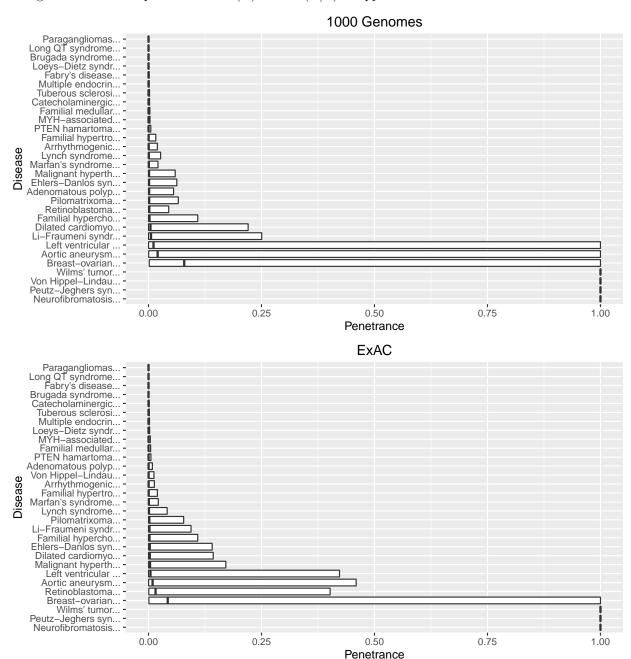
AFR

SAS

3 Penetrance Estimates

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

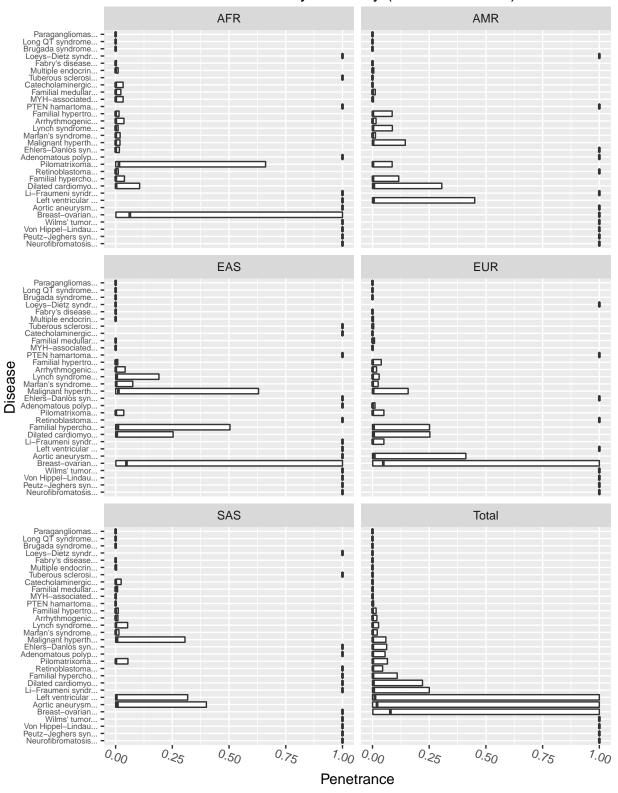
The left end of the boxplot indicates P(D) AND P(V|D) = lower value, the bold line in the middle indicates P(D) AND $P(V|D) = geometric_mean(values)$, the right end of the boxplot indicates P(D) AND P(V|D) = upper value.



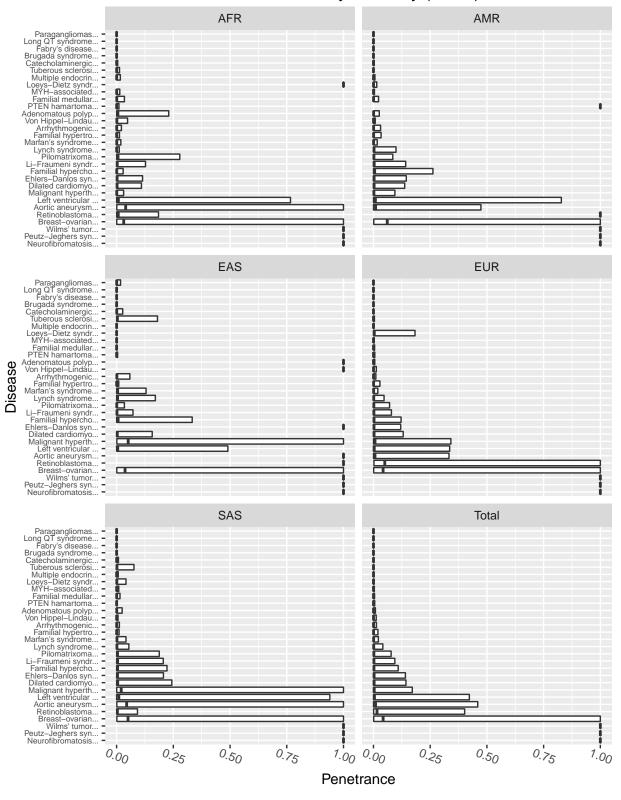
Note: Prevalence ranges of 5x were assumed for all point estimates of prevalence. For example: a point estimate of 0.022 would be given the range 0.01-0.05.

3.2 Penetrance Estimates by Ancestry

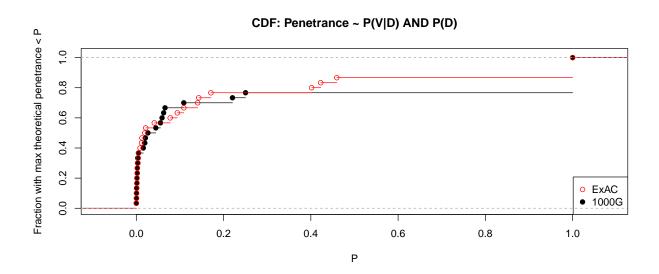
Penetrance by Ancestry (1000 Genomes)



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 Breast-ovarian... Retinoblastoma... Aortic aneurysm... 1e-02 -Left ventricular ... Malignant hyperth... Dilated cardiomyo... Ehlers-Danlos syn... Familial hypercho... Penetrance_ExAC Lynch syndrome... Pilomatrixoma... Li-Fraumeni syndr... Marfan's syndrome... Arrhythmogenic PTEN hamartoma... Familial medullar.. Adenomatous polyp... Multiple endocrin.. Loeys-Dietz syndr... Tuberous sclerosi... Catecholaminergic... Brugada syndrome.. 1e-05 -Fabry's disease... Long QT syndrome... Paragangliomas... 1e-06 1e-04 1e-02 1e-01 1e+00 1e-06 1e-05 1e-03 Penetrance_1000_Genomes

The Pearson correlation is 0.94. Max penetrance values computed using 1000 Genomes are 1.7-fold larger than those computed using ExAC.