

ClinVar Report

James Diao

November 4, 2016

Contents

| | | |
|----------|---|----------|
| 1 | Collect and Merge ClinVar Data | 2 |
| 1.1 | Import ClinVar VCF | 2 |
| 1.2 | Merge ClinVar with 1000 Genomes and ExAC | 2 |
| 2 | Summary Statistics | 3 |
| 2.1 | Fraction of Individuals with Pathogenic Non-Reference Sites | 3 |
| 3 | Penetrance 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 |

Sourcing ClinVar input from: clinvar_2016-10-03.vcf

Sending output to: Report_2016-10-03.pdf

1 Collect and Merge ClinVar Data

1.1 Import ClinVar VCF

Processed ClinVar data frame 117420 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 | 117420 |
| LP/P-ClinVar | 33633 |
| LP/P-ClinVar & ACMG | 6971 |
| LP/P-ClinVar & ACMG & ExAC | 964 |
| LP/P-ClinVar & ACMG & 1000 Genomes | 147 |

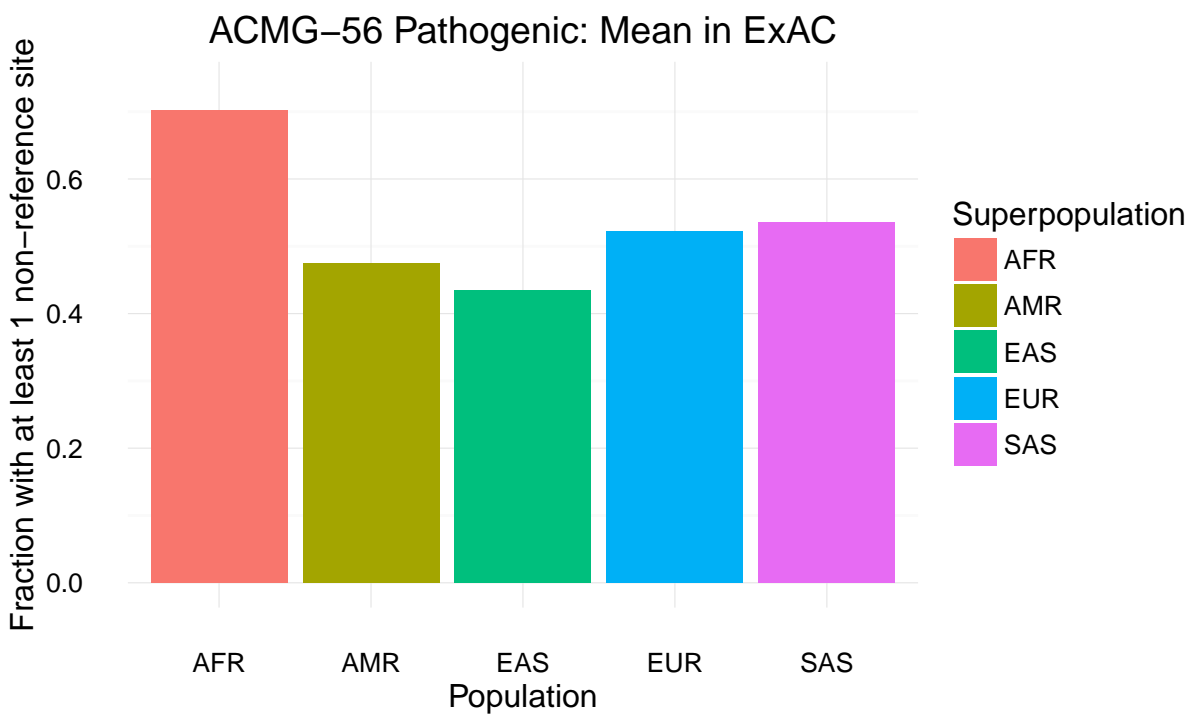
Breakdown of ACMG-1000 Genomes Variants

| 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 |

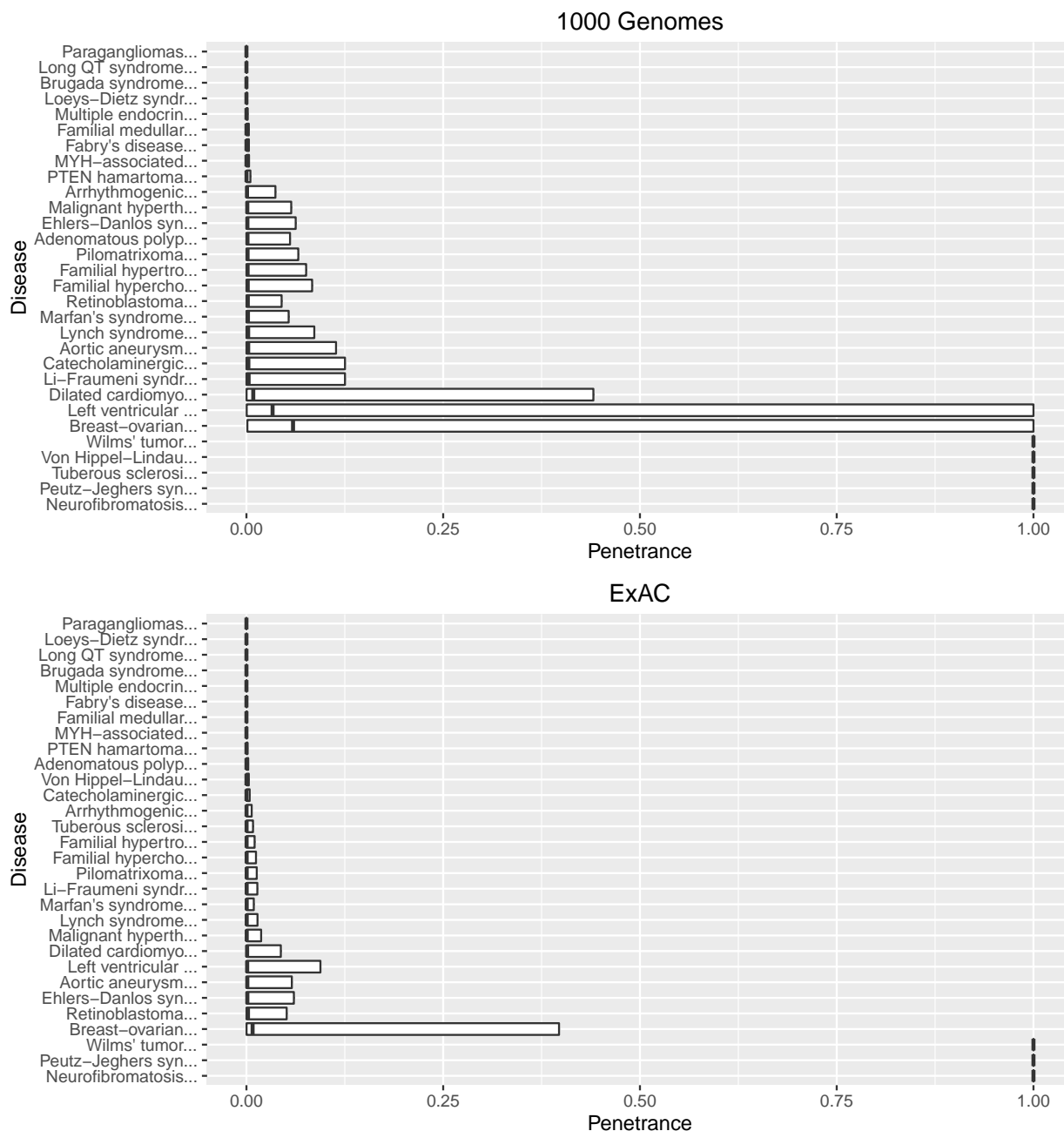
2.1 Fraction of Individuals with Pathogenic Non-Reference Sites



3 Penetrance Estimates

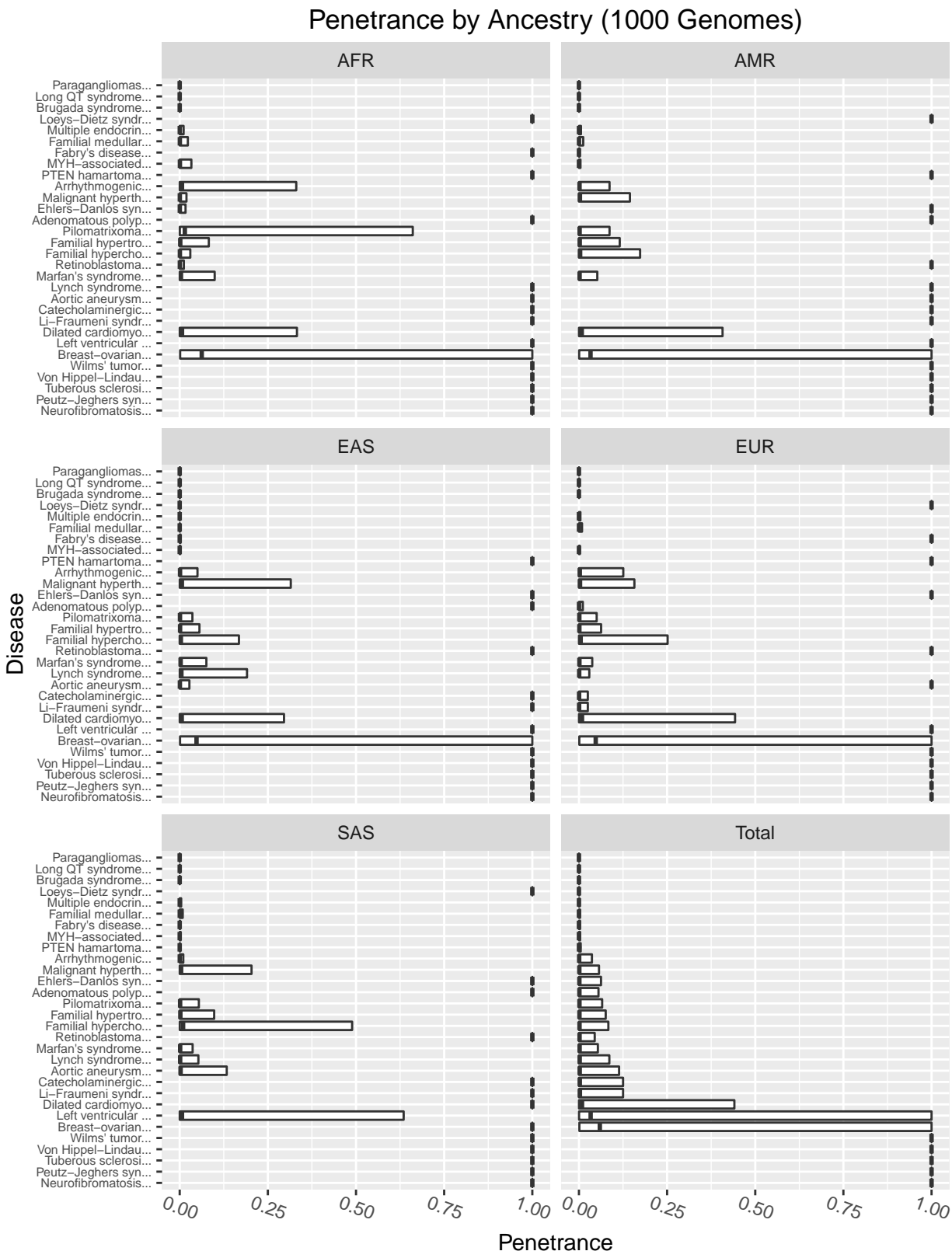
3.1 Max/Min Penetrance as a Function of $P(D)$ and $P(V|D)$

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.

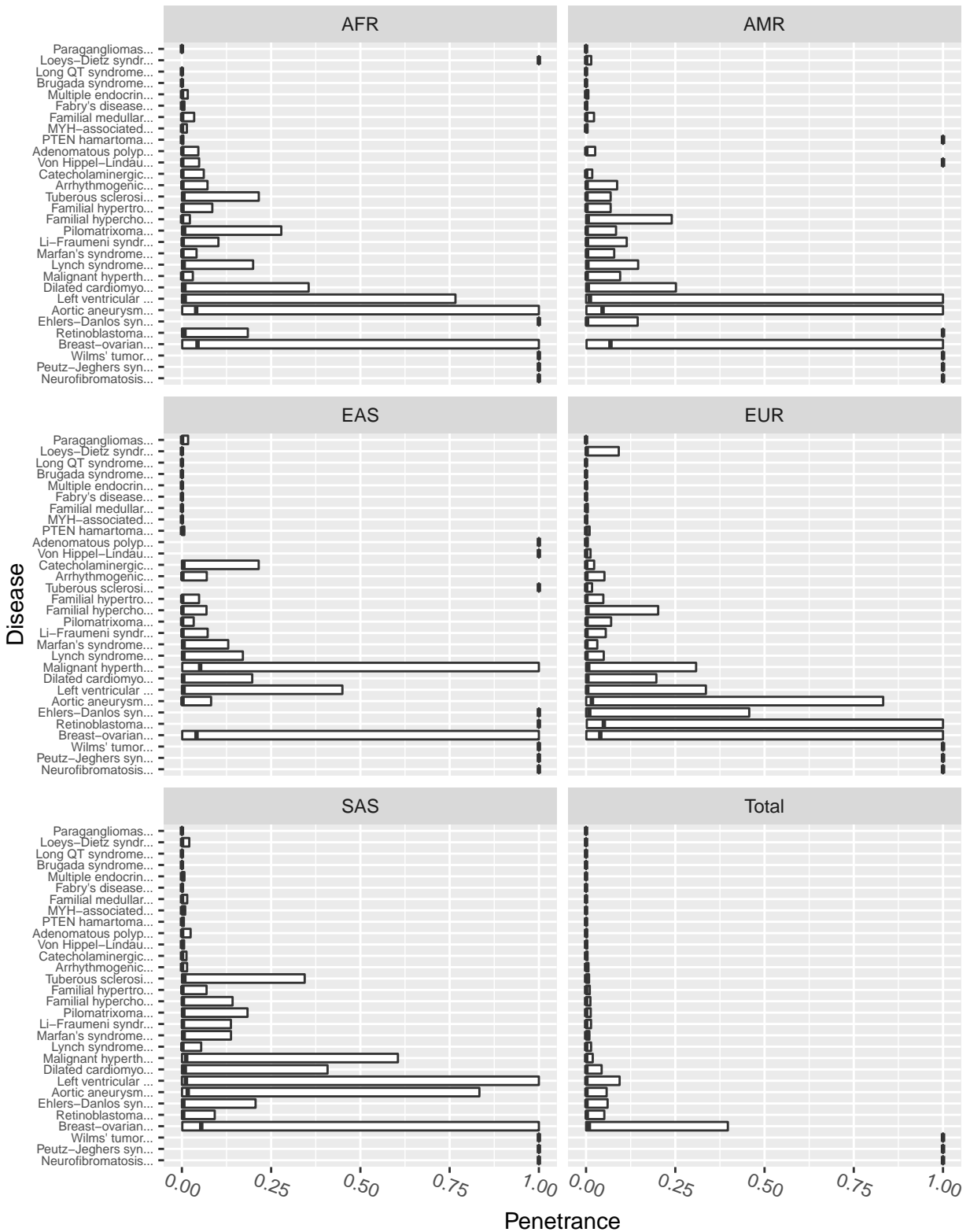


Note: Prevalence ranges of 5x were assumed for all point estimates of prevalence.
For example: a point estimate of 0.022 would be given the range 0.01-0.05.

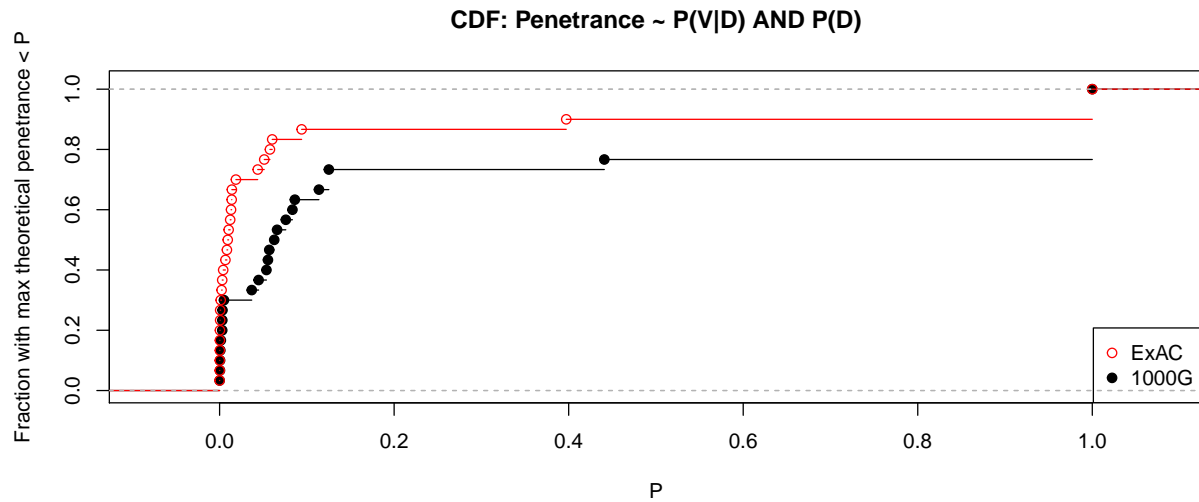
3.2 Penetrance Estimates by Ancestry



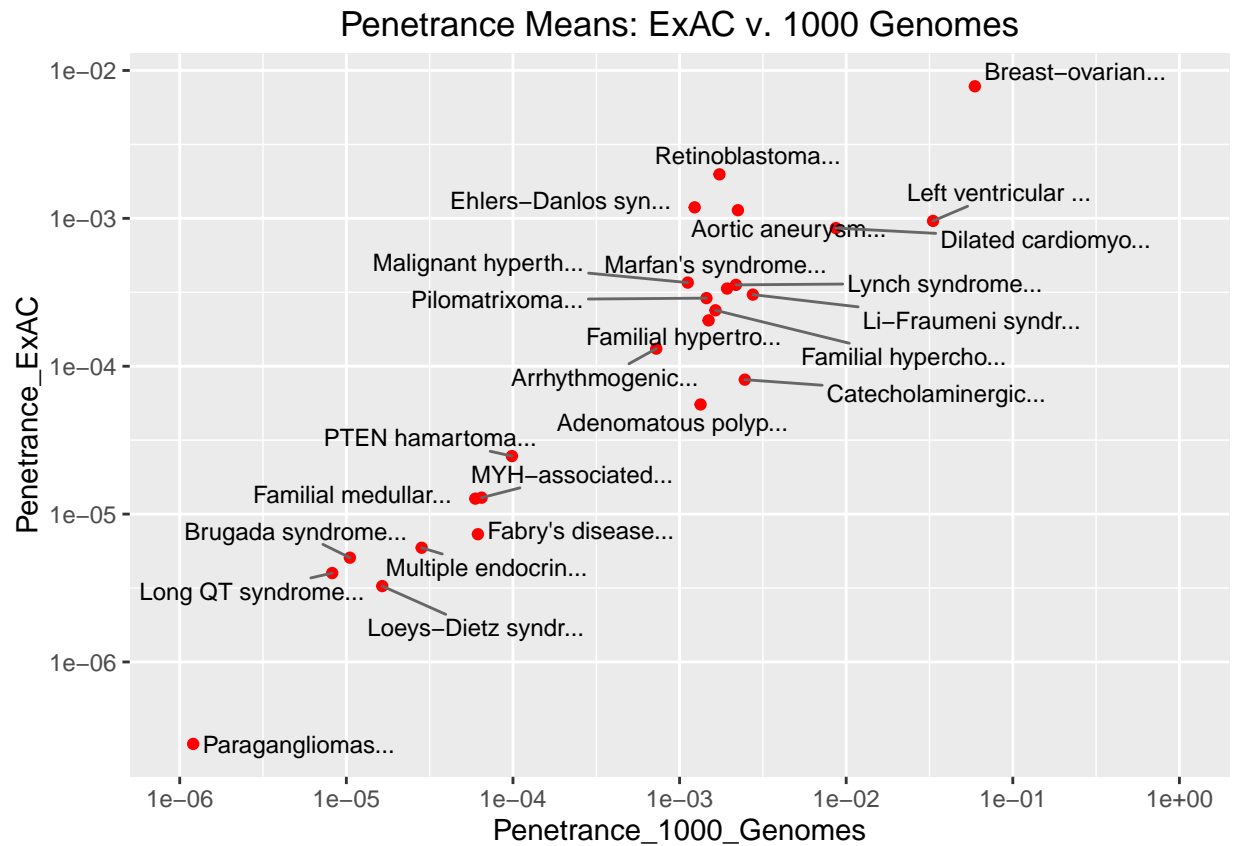
Penetrance by Ancestry (ExAC)



3.3 Empirical CDFs for All Penetrance Plots



3.4 Comparing Mean Penetrance between ExAC and 1000 Genomes



The Pearson correlation is 0.88.

Max penetrance values computed using 1000 Genomes are 7.3-fold larger than those computed using ExAC.