# Presence of Deleterious Variants in the gnomAD Population Data Set

Matt Croken

2022 September 02

# Allele Frequency in Populations

- The compilation and harmonization of genomic sequencing data is an ongoing and critical effort impacting multiple scientific domains
- In the clinical sequencing context, understanding the frequency of alleles in populations is vital to variant interpretation

# The Clinical Context

- Generally, variants occurring more frequently in a population are less likely to be linked to a disease state
- When attempting to detect somatic variants without a 'Normal' control, population-level allele frequencies are used to identify and exclude suspected germline variants

# The Clinical Context, but More Complicated

- There are no established best practices or guidelines for setting an allele frequency threshold
  - Too low risks overwhelming the variant curator
  - Too high risks excluding relevant variants
- As NGS panels trend larger, the risk of error increases with the volume of variants
- Tumor Mutational Burden, an important therapeutic indicator, is usually calculated in a fully automated way

# gnomAD

- The gnomAD database succeeds and builds on many past aggregation efforts
- gnomAD is a carefully curated and nuanced data set
  - It is most frequently used in decidedly un-nuanced ways

# Preliminary Objectives

- Identify the extent to which predicted deleterious variants exist in gnomAD and at what frequencies
- Identify gnomAD variants in OncoKB (cancer domain specific)

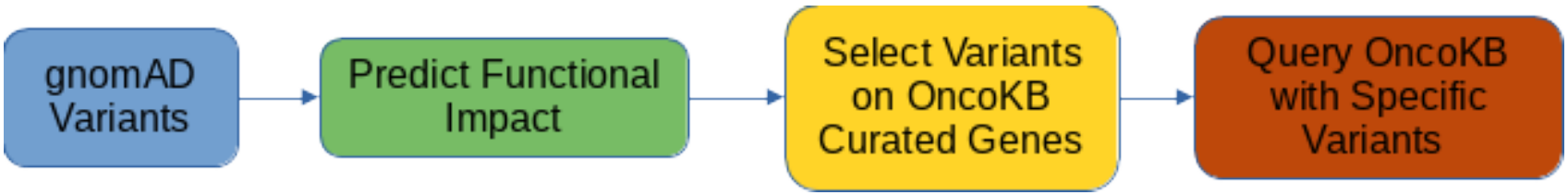
# Methodology - Tools

- github.com/mcroken/pathpop
- bcftools
  - Query and reformat VCF files
- SnpEff
  - Predict effects of genomic variants on transcripts
- GNU Make
  - Workflow orchestration & reproducibility

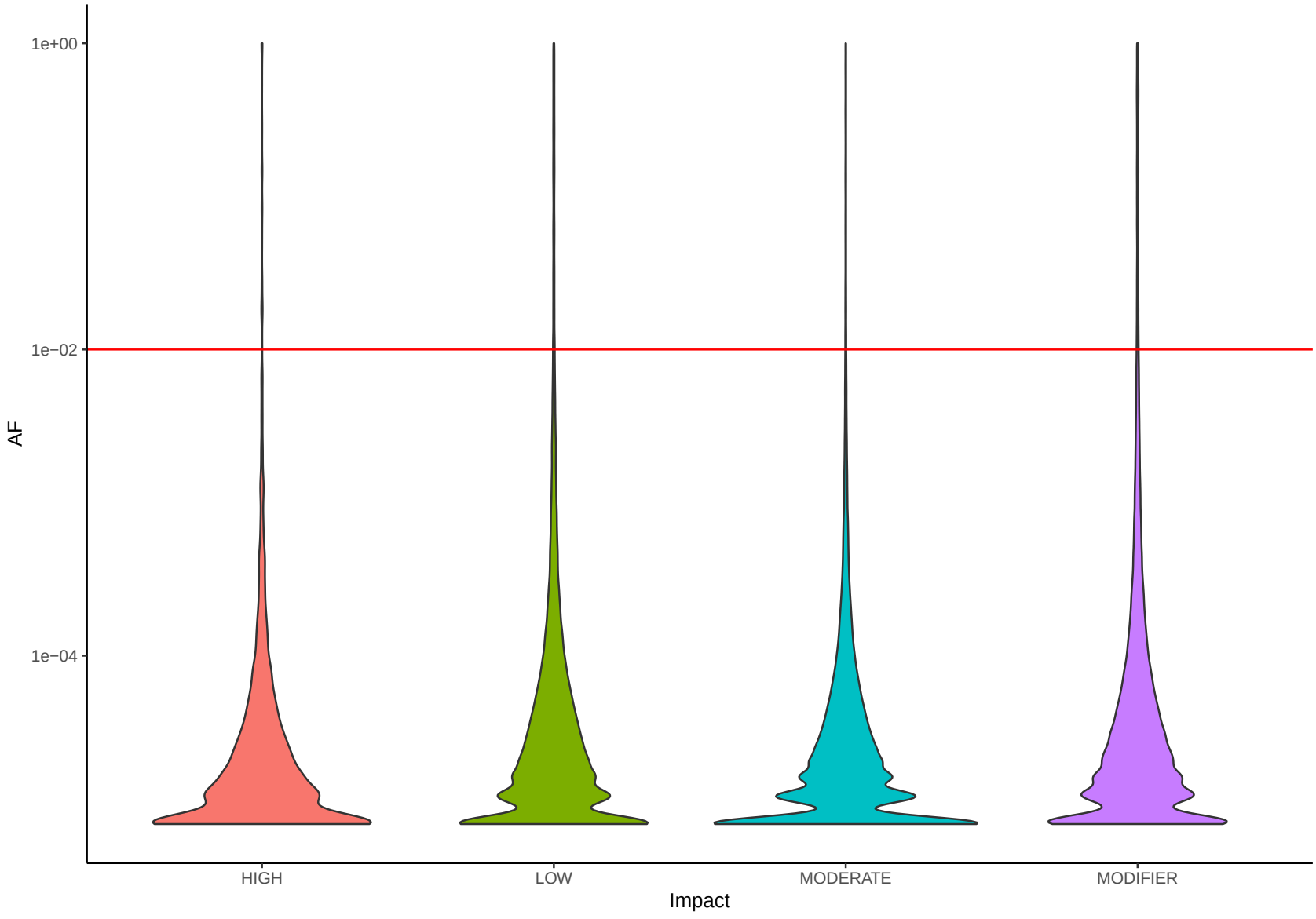
# Methodology - Tools

- OncoKB REST API
  - Identify relevant genes to target
  - Query for oncogenic variants
- Quarto
- Tidyverse
  - Data analysis and visualization

# Analysis Strategy



# Impacts of gnomAD Variants in OncoKB Curated Genes



# Impacts of gnomAD Variants in OncoKB Curated Genes

| Impact | AF greater than 1% | AF less than 1% |
|---|---|---|
| HIGH | 70 | 6114 |
| LOW | 2099 | 93261 |
| MODERATE | 1411 | 125346 |
| MODIFIER | 4446 | 147191 |

# “Germline” (AF $>$ 1%) Variants in OncoKB

| n | Oncogenic Status |
|---|---|
| 11 | Inconclusive |
| 16 | Likely Neutral |
| 156 | Likely Oncogenic |
| 3 | Oncogenic |
| 7761 | Unknown |

# “Germline” (AF $>$ 1%) Variants in OncoKB

| n | Hotspot |
|---|---|
| 6847 | false |
| 1092 | null |
| 8 | true |

# Conclusions

- Limited (but non-zero) number of variants strongly associated with cancer or predicted to be deleterious
- Significant number of variants which are likely oncogenic or deleterious
- As these variants have relatively high allele frequencies, the problem is not easily controlled by raising the AF threshold
- Flag or remove gnomAD variants which are actually reportable in advance

# Future Directions

- Flag or remove gnomAD variants which are actually reportable
  - Production workflow to continually update
- Redo analysis without sub-setting the gnomAD data set
- Query additional databases for cancer-relevant variants
- Periodically query variants excluded as “germline” for novel disease associations.