

# VCFs, Variants, and Python

John Garza

# Recap: Important File Formats

- Sequences
  - FASTA, FASTQ
- Alignments
  - SAM/BAM/CRAM
- Features & Annotations
  - GTF/GFF
  - BED
  - VCF: Variant Call Format

# VCF Annotations & VEP

- Identity: Coordinates & Sequence
- Annotation: Biological & Clinical Significance
- VEP: Variant Effect Predictor
  - Genes and transcripts affected by the variants
  - Regulatory context of the variants
  - Consequence of your variants on the protein sequence
  - SIFT and PolyPhen-2 scores for changes to protein sequence
  - Many more, including custom/user defined

# How VEP Works

- Direct analysis
  - Mutation type from sequence/amino acid change
  - Regulatory context from coordinates
- Database matching
  - ClinVar: clinical significance
  - gnomAD: population frequency
  - COSMIC: somatic mutations in cancer
- Novel Analysis
  - Predict functional impact using computational tools

# Why Annotate?

- Hone in on significant mutations
- Clinical Applications
  - Identify pathogenic mutations
  - Select variants for further investigation
  - Precision medicine
    - Neoantigen targets for immunotherapy
- Research Applications
  - Genotype-phenotype associations
  - Discovery of novel mechanisms

# Demonstration: Manipulating VCFs