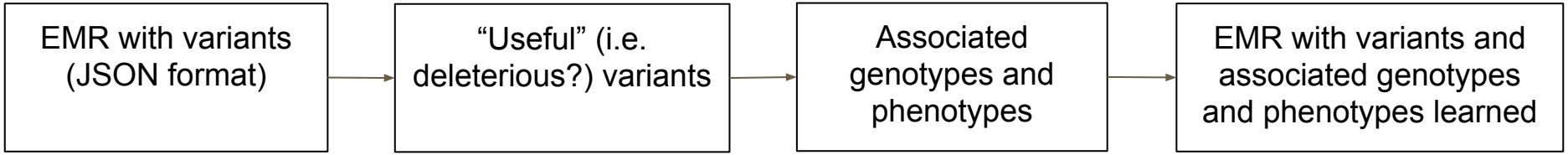
Precision Medicine Toolkit

Hale Kpetigo
Syed Hussain Ather
Rob Guthridge
Blythe Hospelhorn
Luli Zou

Motivation

- Collaboration between clinical work and bioinformatics is difficult because EMR data exchange format are different.
- The industry has been building large databases of bioinformatic data that could be useful for clinicians when diagnosing patients
- Some of these databases are:
  - TOPMed (Trans-Omics for Precision Medicine)
  - dbGaP
  - SRA
  - GWAS (Genome-Wide Association Study)
- Doing searches in these databases is currently manual and tedious

Precision Medicine Toolkit



Use case: While diagnosing a client who has a mutation on a specific gene, can I search public (or restricted access) database for the same variant, or populations of similar patients?

Variant Annotation Tools

Annovar

Input: Annovar formatted text table, OR VCF file

Output: Input table with additional columns, and/or VCF file with annotations as INFO fields.

SNPEff

Input: chromosome, position, ref. allele, alt. allele OR vcf

Output: JSON file with alternate allele frequency

Genome Reference Databases

TOPMed (via Bravo API)

Input: chromosome, position, ref. allele, alt. allele OR vcf

Output: JSON file with alternate allele frequency

GWAS

Input: GWAS catalog (a .tsv file, each line is a variant)

Output: Name of study/disease trait, pubmed link

Sequence Rate Archives

SRA (SRR)

Input: SRA files from SRA database
SRA -> fastq -> bam -> VCF
Output: VCF with DNA polymorphism data

dbGaP

Product Roadmap

- Complete first iteration of pipeline with simple input (one vcf with one chromosome information)
- Features to add in the future could include:
  - Input with multiple chromosomes in json
  - Expose tool to the web so that pipeline can be called via webapi
  - Expose tool to the web and offer visualization of outputs
  - Create docker components for the pipeline
  - Ability to search private databases