Work in Progress

July 26 2018 – Jasper Ho

HIV-1 Resistance Prediction

- Guides the clinical management of HIV-1 infection
- Predicts resistance of *in vivo* HIV-1 to specific ARVs
- Two general types:
  - Phenotypic
  - Genotypic
- Genotypic can be:
  - Rules-based
  - Data-driven

	Phenotypic	Genotypic
Empirical?	Yes	No
Speed	Slow	Fast
Cost	$$	$

HIVdb – What Is It?

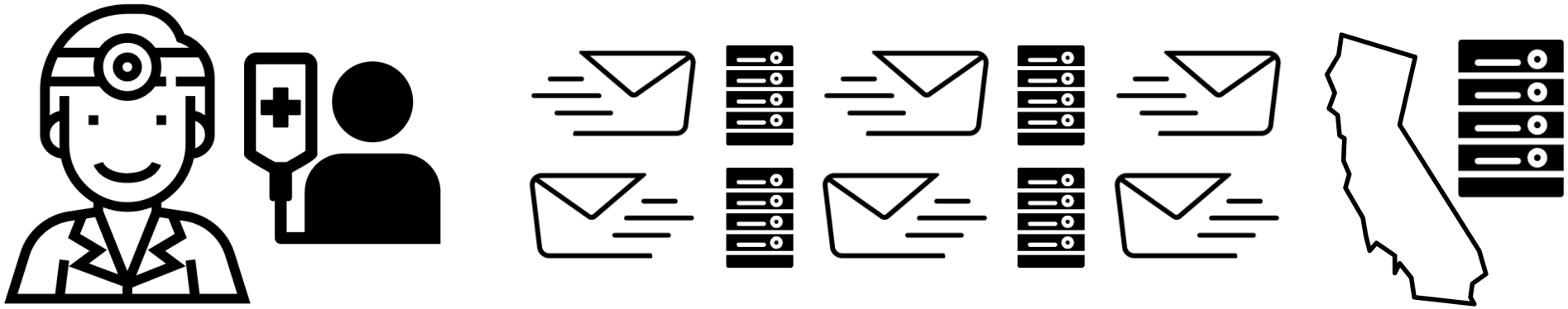
- Rules-based **genotypic resistance interpretation system**
- Stanford University
- Complementary to the Stanford HIV Drug Resistance Database
- Resistance level predictions from nucleotide or amino acid HIV-1 *pol* sequences
- Accessible through the Sierra Web Service, or a web interface

The Algorithm

1. Alignment of query sequence to subtype B reference *pol* sequence
2. Mutations identified in aligned sequence
3. Drug resistance mutations (DRMs) identified
4. Penalty scores of each DRM present are summed for each ARV
5. ARV's total penalty score is converted to 1 of 5 resistance levels: susceptible, potential low-level resistance, low-level resistance, intermediate resistance, high-level resistance

HIV-1 sequence data is **highly sensitive patient information**

confidentiality, stigma and legal implications of source attribution



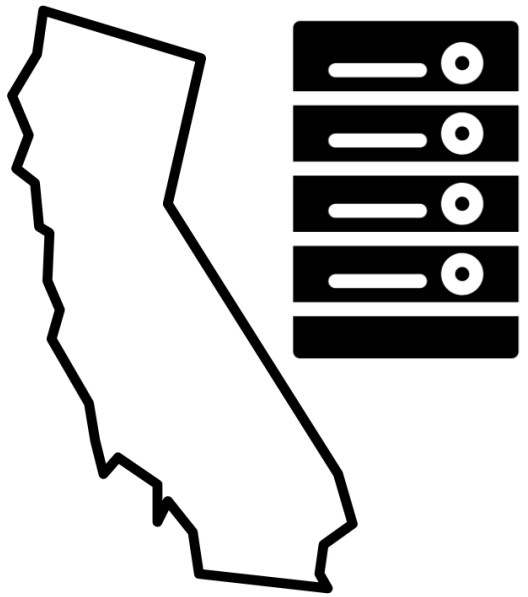
what's happening on the server?

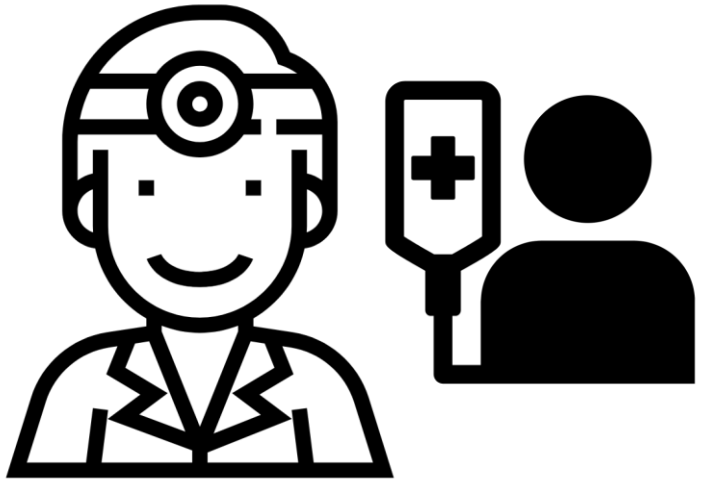
algorithm version

source code

data retention

server uptime





how can health providers be impacted?

network outages

software versions

sierra-local

- A software package for the implementation of the HIVdb algorithm
  - Python
  - Completely local
    - Reliable
    - Secure for patient information
  - Algorithm version aware
  - Fewer dependencies
  - Faster

Algorithm Specification Interface (ASI)

Betts and Shafer, 2003.
doi: 10.1128/JCM.41.6.2792-2794.2003

```
<DRUG>
    <NAME>RAL</NAME>
    <FULLNAME>raltegravir</FULLNAME>
    <RULE>
        <CONDITION>
            <![CDATA[
            SCORE FROM(51Y => 15, MAX ( 66A => 15, 66I => 15, 66K => 60 ), MAX ( 92G => 15,
            92Q => 30, 92V => 30 ), 95K => 10, 97A => 10, 118R => 30, 121Y => 60, MAX ( 138A
            => 15, 138K => 15, 138T => 15 ), MAX ( 140A => 30, 140C => 30, 140S => 30 ), MAX
            ( 143A => 60, 143C => 60, 143G => 60, 143H => 60, 143K => 60, 143R => 60, 143S =>
            60 ), MAX ( 148H => 60, 148K => 60, 148N => 10, 148R => 60 ), MAX ( 151A => 15,
            151L => 30 ), MAX ( 155H => 60, 155S => 30, 155T => 30 ), 157Q => 10, MAX ( 163K
            => 15, 163R => 15 ), 230R => 20, 263K => 25, (138AKT AND 140ACS) => 15, (74FM AND
            148HKR) => 10)
            ]]>
        </CONDITION>
        <ACTIONS>
            <SCORERANGE>
            <USE_GLOBALRANGE/>
            </SCORERANGE>
        </ACTIONS>
    </RULE>
</DRUG>
```

Overview

- Pre-processing
  - Alignment to subtype B reference sequence
    - NucAmino only returns a list of mutations and not the aligned query sequence itself
  - Gene identification
  - Mutation site classification
  - Sequence trimming
    - Leading and trailing regions with >30% ‘low quality’ sites
      - Highly ambiguous, stop codons, unusual mutations, frameshifts, APOBEC-mediated G-to-A hypermutation
  - Subtyping (in progress)
- Resistance scoring
  - Drug-by-drug condition checking
  - Penalty scores added
- Output

Comparing against HIVdb Sierra

- Concordance
- Speed / performance

Dataset

- Stanford HIV Data Base's **genotype-treatment correlation dataset**
  - Nucleotide sequence, subtype, date and region of collection, list of ARVs exposed to *in vivo*
  - 112,724 RT
  - 105,694 PR
  - 12,332 IN
  - 230,750 total
- Randomly partitioned into same-gene 1000-sequence files

Concordance

- Processed entire dataset with both sierra-local and SierraPy
- 230,719 / 230,750 queries have identical scores (99.987%)
- 31 queries do not have identical scores
  - 5 protease
  - 22 reverse-transcriptase
  - 2 integrase
- Causes:
  - Incomplete validation steps
  - Trimming differences
  - Refusal to process; no genes found

Speed / Performance

- Random selection without replacement of 10 files per gene (30 total)
  - ~10,000 sequences across 10 files per gene
- Submitted files to both *sierra-local* and SierraPy

Mean processing speed, sequences per second

	sierra-local	SierraPy
Protease	112.17	16.01
Reverse-transcriptase	45.36	6.12
Integrase	37.36	5.19

7x faster

Next Steps

- Finishing validation and pre-processing
- Manuscript preparation
- Preprint submission