HIVDB Genotypic Drug Resistance Interpretation Program

Robert Shafer, MD
Professor of Medicine
Stanford University

Disclosures

- Gilead Sciences (2022): Advisory board and speaking honorarium.
- ViiV Healthcare (2022): Speaking honorarium.

These are my disclosures.

Goals of Genotypic HIVDR Testing

- Individual patient management.
- Surveillance of transmitted and acquired HIVDR.
- Inclusion criteria and outcome variables in clinical trials.

1. Genotypic HIVDR testing is performed for three main purposes.
2. <> In UICs it is performed to optimize the management of individual patients.
3. <>In LMICs, it is performed primarily for the surveillance of transmitted and acquired HIVDR.
4. <>In clinical trials, it is used as inclusion criteria and as an outcome variable.
5. For the 2nd and 3rd goals, it has become important to use a consistent approach so that the results of different studies can be compared with one another and can be combined.

Limitations for Individual Patient Management

- Cannot recommend specific ART regimens
  - Does not consider additional clinical data such as past ART history, past genotypic resistance data, virus load, CD4 count, and potential toxicities.
  - Users must also be knowledgeable about the principles of HIV treatment and published treatment guidelines.

1. For the 1st goal, the interpretation program is intended to be educational, and it has several limitations for individual patient management.
2. <>The program lacks the ability to recommend specific regimens for two main reasons
3. <>First, it does not consider additional clinical data such as past ART history, past genotypic resistance data, virus load, CD4 count, and potential toxicities.
4. <>Second, it does not consider the implicit knowledge that most care providers have and that is often explicitly outlined in published treatment guidelines.

Implementation

- Rules based on individual and combination mutation penalty scores.
- Comments that accompany mutations.

1. The interpretation system comprises numerical penalties for individual DRMs as well as additional penalties when certain mutations occur in combination.
2. The penalties associated with individual DRMs are added to the penalties associated with DRM combinations to derive a total penalty for the drug.
3. Each DRM is accompanied by a comment that contains information on the prevalence of the mutation, how it influences susceptibility to drugs belonging to the relevant drug class, and whether there are data on how it might influence the response to therapy.
4. A rules-based interpretation system seems somewhat archaic in this era of machine learning and AI.
5. However, for reasons that go beyond the scope of this talk, we don't believe that it is currently possible to develop an optimal ML algorithm for genotypic HIVDR interpretation.

HIVDB Program: Levels of HIVDR

Resistance Level	Definition	Score range
Susceptible	No evidence of reduced susceptibility	<10
Potential low-level resistance	DRMs consistent with previous ARV exposure or DRMs associated with resistance only when they occur with other DRMs	10-14
Low-level resistance	DRMs associated with a reduction in vitro ARV susceptibility or a suboptimal virological response to ARV treatment.	15-29
Intermediate resistance	A high likelihood that ARV activity would be reduced. However, the ARV would likely still retain antiviral activity.	30-59
High-level resistance	A level of resistance similar to that observed in viruses with the highest levels of reduced in vitro susceptibility or in viruses that have little or no virological response to ARV treatment.	≥60

1. Individual DRMs receive penalties that range from 5 to 60 for each drug belonging to a drug class.
2. The DRM penalties are designed such that when the individual and combination penalties for a drug are added together, they generally fall in a range from 0 to slightly more than 100.
3. This table indicates how the total penalty score for a drug is translated into one of 5 different resistance levels ranging from susceptible to high-level resistance and what is meant by each level.
4. The table also indicates what is meant with the different levels of reduced susceptibility.

5. "Susceptible" is assigned when a virus displays no genotypic evidence of reduced susceptibility compared with a wild-type virus.
6. "Potential low-level resistance" is assigned when a virus has DRMs consistent with previous ARV exposure or contains DRMs associated with resistance only when they occur with other DRMs.
7. "Low-level resistance" is assigned when a virus has DRMs associated with a reduction in in vitro ARV susceptibility or a suboptimal virological response to ARV treatment.
8. "Intermediate resistance" is assigned when there is a high likelihood that

ARV'activity would be reduced, but that the drug would still retain significant antiviral activity.
9. "High-level resistance" is assigned when the predicted level of resistance is similar to that in viruses with the highest levels of reduced in vitro susceptibility or in viruses that would be expected to have little or no virological response to treatment with the indicated drug.
10. A few DRMs have negative scores for certain drugs. These hyper-susceptibility mutations can mitigate the effect of other DRMs on that drug.

How DRM Scores are Derived

- Selective drug pressure
- Reduction in in vitro susceptibility ("phenotypic data")
- Reduce virological response to a new ART regimen
- Expert opinion

1. The DRM penalty scores and comments are created and updated based on based on 3 main types of data/considerations:
2. <>First, are the mutations selected by the drug.
3. <>Second, do the mutations reduce drug susceptibility in vitro.
4. <>Third, is there any evidence that the mutations reduce the virological response to a regimen containing the drug for which the scores are developed.
5. <>Finally, for mutations for which there is insufficient available data, consensus expert opinion is also considered.
6. In the following three slides, I will drill down into each type of data.

7. Now lets drill down into each type of data.
8. First selective drug pressure: Generally, mutations that arise naturally – also referred to as polymorphisms – are generally not assigned penalties or assigned very low penalties.
9. Empirically such mutations have been very rarely associated with reduced drug susceptibility which is fortunate and reflects the fact that for the main classes of drugs resistance does not arise naturally
10. Mutations that are selected in vitro by a drug are often the main mutations that arise in patients, but this is not always the case.

11. Some mutations selected in vitro never occur in patients and frequently mutations that have not been reported in vitro do arise in patients with VF on a drug.
12. Second, in vitro drug susceptibility or phenotypic data: Susceptibility data can be performed on site-directed mutants. In this scenario it is possible to strongly link the effects of individual mutations or combinations of mutations on reduced drug susceptibility.
13. However, most DRMS and combinations of DRMs have not been studied in this way. Therefore, much of the in vitro susceptibility data comes from clinical isolates. These data are more complex because clinical isolates often contain many different combinations of DRMs and sometimes mutations that are not considered DRMs – often referred to as backbone mutations – can modulate the effect of a DRM combination.
14. Third, virological response to a regimen containing the drug of interest: Such orrelations between genotype and virological suppression are relevant because sustained virological suppression is the main goal of ARV therapy.
15. There have been several highly informative clinical trials that have demonstrated associations between specific pre-therapy DRMs and the risk of VF.
16. However, many retrospective studies have had too few patients relative to the large number of covariates associated with response to therapy – for example, the patient's complete ART history, plasma VL, the drugs used in combination with the drug of interest, and the level of adherence to therapy.
17. Moreover, most retrospective studies have been confounded by the fact that the results of GRT were used to guide the choice of therapy. These data are obtained from the published literature.

Deriving DRM Scores: Selective Drug Pressure

- Is the mutation selected by the drug in vitro?
- Is the mutation selected by the drug in patients?

Major DRMs rarely occur in the absence of selective drug pressure. However, many accessory DRMs are often polymorphic.

1. <>When a drug is first found to have ARV activity, in vitro passage experiments are performed to identify the mutations that arise under selective drug pressure.
2. These mutations are compared with those associated with other drugs belonging to the same class and are assessed for drug susceptibility and replication fitness.
3. <>Once the drug is used in patients, it becomes possible to identify the mutations that arise under in vivo circumstances.
4. The mutations that arise in patients are typically greater in number than those that arise during in vitro passage experiments.
5. However, some of the mutations that arise in vitro never arise in patients presumably because they are not sufficiently replication competent.
6. <>Major DRMs rarely occur in the absence of selective drug pressure. However, many accessory DRMs are often polymorphic.

Deriving DRM Scores: In Vitro Susceptibility Data

- Site-directed mutants
- Clinical isolates

1. In vitro susceptibility data – also referred to as phenotypic data – can be performed on site-directed mutants and on clinical isolates.
2. Testing of site-directed mutants is often done early during drug development. When viruses are created that contain just a single DRM or a small number of DRMs, it is possibly to reliably assess the influence of these DRMs on drug susceptibility.
3. However, most DRMS and combinations of DRMs have not been studied in this way and most published in vitro susceptibility data comes from clinical isolates.
4. In vitro susceptibility data on clinical isolates are more difficult to interpret because clinical isolates often contain many different combinations of DRMs and sometimes mutations that are not considered DRMs – often referred to as backbone mutations – can modulate the effect of a DRM combination.

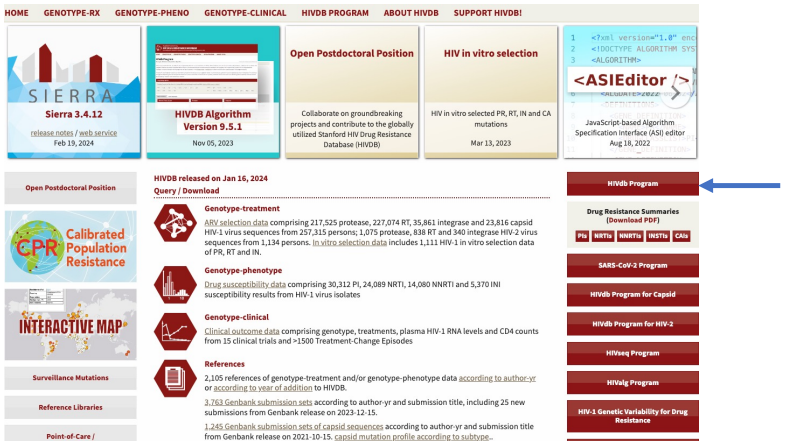
Deriving DRM Scores: Virological Response

- Clinical trials
- Retrospective cohort studies

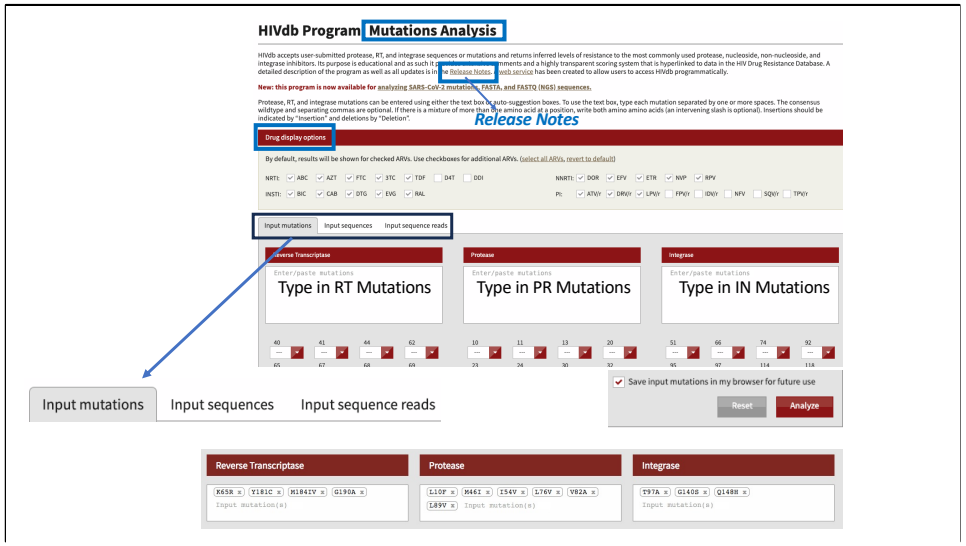
1. The third type of data that we use to derive DRM scores is clinical data.
2. This involves determining whether the presence of a mutation prior to starting a regimen containing a specific drug reduces the virological response to the regimen.
3. One might expect this to be the most important type of data for genotypic resistance interpretation.
4. However, in practice this is not the case for the following reasons.
5. First, such analyses are complicated by the presence of multiple different patterns of baseline mutations prior to starting a new regimen, the presence of multiple covariates including past ART history, plasma VL, adherence to therapy, and by the drugs used in combination with the drug of interest.
6. Second, the raw data of this type have rarely been made available for meta-analyses.
7. Nonetheless, there have been several highly informative clinical trials that have demonstrated associations between specific pre-therapy DRMs and the risk of VF.
8. However, many retrospective studies have had too few patients relative to the large number of covariates associated with response to therapy – for example, the patient's complete ART history, plasma VL, the drugs used in combination with the drug of interest, and the level of adherence to therapy.

9. Moreover, most retrospective studies have been confounded by the fact that the results of GRT were used to guide the choice of therapy.

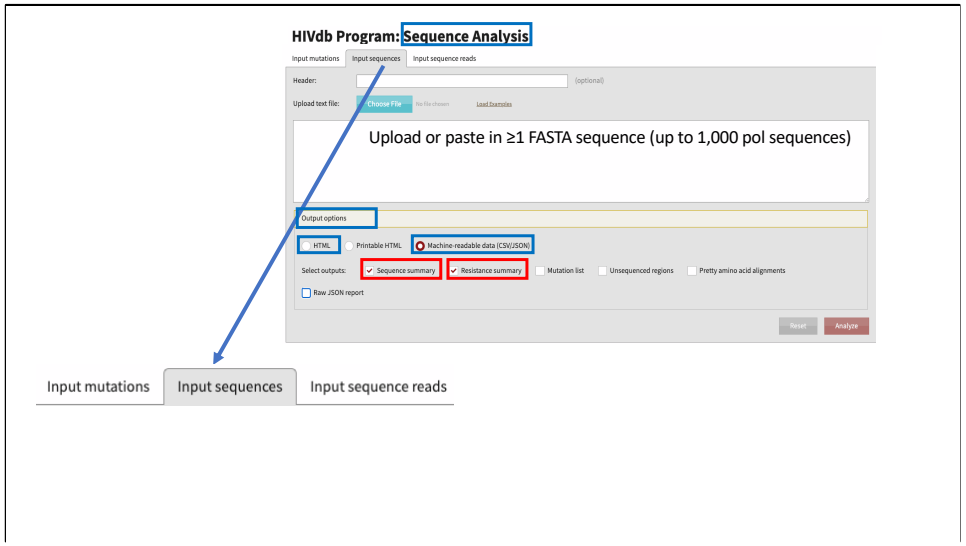
HIVDB Home Page => HIVDB Program



1. The second half of this talk will review the use of the Stanford database genotypic resistance interpretation program.

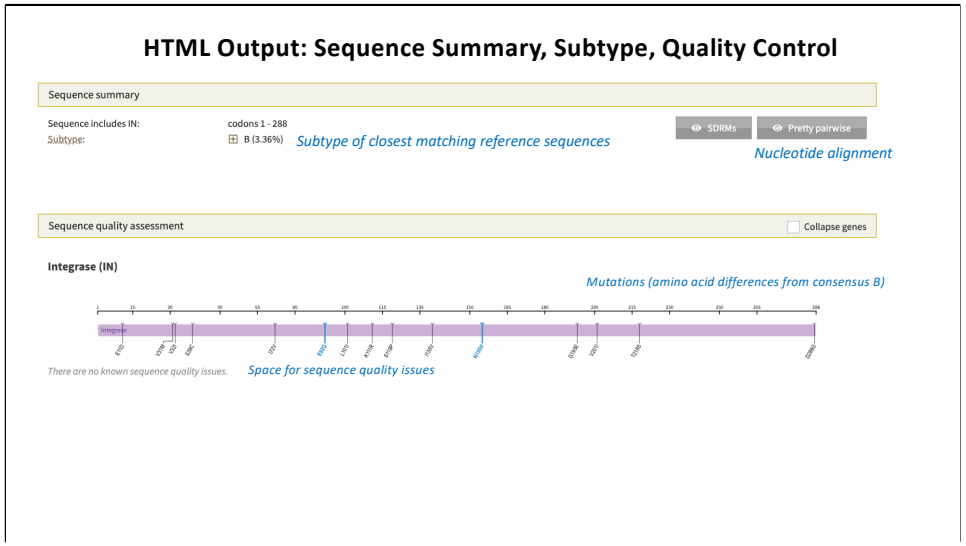


1. <>The program accepts lists of mutations, one or more FASTA sequences, or a FASTQ file.
2. FASTA sequences usually represent Sanger sequences or the consensus of the many reads from an NGS platform.
3. This slide shows how lists of mutation are analyzed.
4. I've highlighted several additional parts of the page including the fact that the user selected the tab "Input mutations, <> a link to the Release Notes; and <> the Drug Display Options. By default, the program shows only those ARVs that are still being used.
5. <> Most often users type in RT, PR, and/or IN mutations in the three text boxes. However, it is also possible to use drop down menus to select each of the DRMs.
6. <>The text boxes allow the mutations to be entered in a variety of different formats and the entered mutations are displayed in a consistent manner.



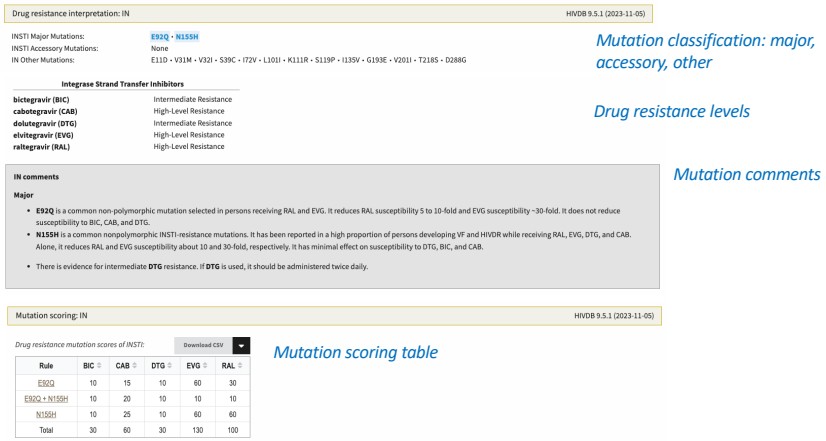
1. <>This slide shows how to either paste in or upload a list of FASTA sequences.
2. The program can handle up to about 1000 complete pol FASTA sequences.
3. <>The slide also shows that the two main output options are HTML or a machine-readable file – usually csv file.
4. <>Among the machine readable formats, the most useful are the Sequence summary and the Resistance Summary.
5. The HTML output is useful if a small number of sequences is entered but csv files are much more useful if many sequences are entered.

HTML Output: Sequence Summary, Subtype, Quality Control



1. This slide and the next one will review the HTML output of the interpretation program.
2. The top part of the result shows the region that was sequenced, the closest matching reference sequence, its genetic distance from the submitted sequence, and a link to the nucleotide alignment.
3. This example shows that the sequence contained the complete integrase gene.
4. By clicking on the plus sign, the top 10 closest matching reference sequences and their subtypes can be viewed
5. <>The next part of the output contains a figure showing the sequence's mutations defined as amino acid differences from the consensus B reference sequence.
6. Drug resistance mutations and mutations that likely reflect sequence quality control issues are indicated by different colors.
7. Beneath the figure there is a textual summary of sequence quality control issues if any are present. These include stop codons, frame shifts, mutations suggesting that APOBEC-mediated G-to-A hypermutation is present, and other unusual mutations.

HTML Drug Resistance Interpretation



Mutation classification: major, accessory, other

Drug resistance levels

Mutation comments

Mutation scoring table

1. For integrase, protease, and capsid, the mutations are divided into major, accessory, and other where those in other do not receive mutation penalty scores.
2. For RT, the mutations are divided into NRTI, NNRTI, and other.
3. <>The next part of the output contains the drug-resistance interpretation based on the five levels previously reviewed on an earlier slide and a list of mutation comments.
4. <>The final part of the output for each gene contains the mutation scoring table, which shows how the individual and combination DRM penalties are added up to yield the total score that is then used to assign one of the 5 drug levels.

Sequence Summary CSV File

Sequence ID	Taken from sequence header
Genes	PR, RT, and/or IN
Start & stop positions	For PR, RT, and IN (6 columns)
Subtype (%)	Closest matching subtype (% nucleotide distance)
Mixture rate (%)	Proportion of positions
PR mutations	Major, Accessory, Other (3 columns)
RT mutations	NRTI, NNRTI, Other (3 columns)
IN mutations	Major, Accessory, Other (3 columns)
SDRMs	Surveillance DRMs: PI, NRTI, NNRTI, INSTI (4 columns)
TSMs	Treatment-selected mutations: PI, NRTI, NNRTI, INSTI (4 columns)
Quality control issues	Frame shifts, Indels, Stop codons, APOBEC, Unusual mutations (12 columns)
Permanent link	URL containing each of the mutations

1. As noted earlier, the spreadsheet output options are recommended for research studies in which many sequences are submitted. This table will review the output of the Sequence Summary file.
2. Overall, the file contains one row for each sequence and 40 columns.
3. <><>The first column contains the unique Sequence ID and the next 7 columns indicate which genes were sequenced and the range of positions that were sequenced.
4. <>There is a column with the genetic distance to the closest matching subtype and a column indicating the proportion of positions containing a mixture of more more than one nucleotide.
5. In newly diagnosed untreated persons, the proportion of positions with mixtures has been shown to correlate with the duration of infection.
6. <>The next 9 columns list the 3 categories of mutations for PR, RT, and IN.
7. <>The next 4 columns list the DRMs which have been used for surveillance of transmitted drug resistance.
8. The surveillance DRMs or SDRMs overlap a lot with the DRMs shown in the preceding columns. However, they differ in that none are polymorphic and that none are extremely rare.
9. <>The treatment-selected mutations or TSMs also overlap with the DRMs shown

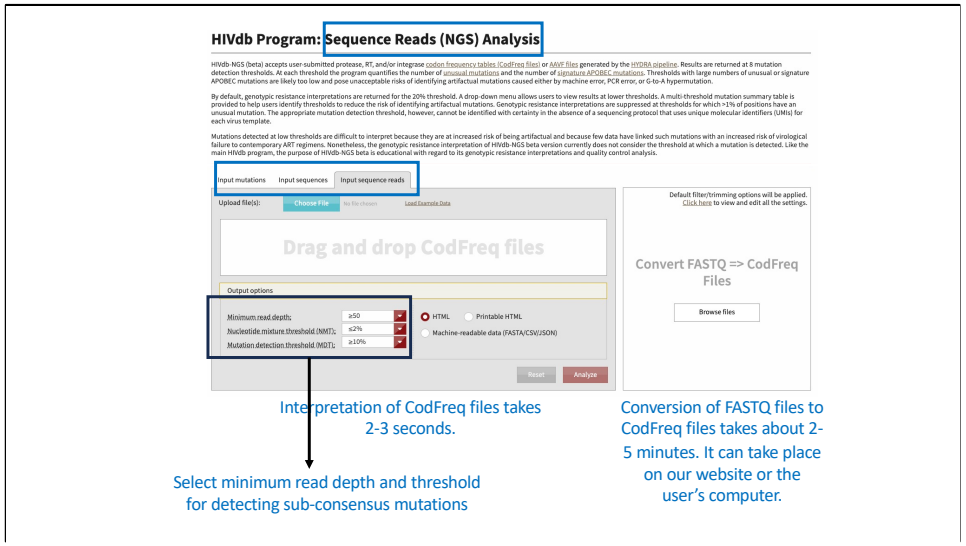
in the preceding columns. In addition to the well-established DRMs, they include nonpolymorphic mutations that are significantly more common in patients receiving ART but which have not been well studied often because they are rare..

10. <>There are 12 columns highlighting QC issues.
11. <>The last column is a URL containing each of the mutations which allows the HTML output to be viewed.

Resistance Summary CSV File

Sequence ID	Taken from sequence header
Genes	PR, RT, and/or IN
PR mutations	Major, Accessory, Other (3 columns)
RT mutations	NRTI, NNRTI, Other (3 columns)
IN mutations	Major, Accessory, Other (3 columns)
Drug scores	Sum of mutation penalty scores
Drug levels	1: susceptible, 2: potential low-level, 3: low-level, 4: intermediate, 5: high-level
Algorithm name	HIVDB
Algorithm version	Version number
Algorithm date	Date algorithm last updated

1. This slide summarizes the columns shown in the Resistance Summary spreadsheet.
2. <>In addition to having the 9 columns showing each of the PR, RT, and IN mutations, it also has columns showing total score and resistance levels for the drugs belonging to each drug class.



1. I will now turn to analyzing NGS FASTQ sequence files.
2. This requires clicking on the "Input sequence reads tab" .
3. Analyzing FASTQ sequences requires two steps.
4. <>The first requires converting one or more FASTQ files to a format we developed called a codon frequency or CodFreq file. The following slide will describe CodFreq files in detail.
5. If two FASTQ files have the same prefix, they are considered to be paired.
6. The conversion of a FASTQ file to a CodFreq file takes about 2-5 minutes depending on the size of the file. The conversion can be done either on our website or on the user's computer.
7. <>One or more CodFreq files can then be uploaded to the interpretation program for rapid analysis.
8. <>The user can then specify three parameters: (1) the minimum number of reads required for a region to be considered sequenced, (2) the maximum allowable proportion of mixed nucleotides across the entire sequence, and (3) the mutation detection threshold which is the minimum proportion of reads containing a mutation for it to be considered present provided the threshold does not result in the maximum allowable proportion of mixed nucleotides to be exceeded.
9. By default the minimum number of reads is set at 50, the nucleotide mixture

threshold is set to 2%, and the mutation detection threshold is set to 10%.

CodFreq files

Make it possible to rapidly evaluate NGS data.

Provide an important measure of quality control - level of background noise.

Prevent linkage of mutations.

Several advantages compared with variant call format (VCF) files:

- Interpretable without a reference sequence.
- Can be used independently from accompanying SAM/BAM file.

Gene	Pos	NumReads	Codon	CodonReads	AA	Pcnt	
RT	200	34068	ACA	26636	T	0.782	
RT	200	34068	GCA	7116	A	0.209	→ Sub-consensus variant
RT	200	34068	ACT	105	T	0.003	
RT	200	34068	ACG	102	T	0.003	
RT	200	34068	GCG	30	A	0.001	
RT	200	34068	ATA	18	I	0.001	} Background noise
RT	200	34068	GTA	11	V	0	
RT	200	34068	ACAA	10	T	0	
RT	200	34068	GCAA	9	A	0	
RT	200	34068	TCA	9	S	0	
RT	200	34068	ACC	5	T	0	
RT	200	34068	CCA	5	P	0	
RT	201	34693	AAA	33699	K	0.971	
RT	201	34693	AAG	679	K	0.02	
RT	201	34693	GAA	131	E	0.004	
RT	201	34693	AGA	129	R	0.004	} Background noise
RT	201	34693	ATA	12	I	0	
RT	201	34693	TAA	12	*	0	
RT	201	34693	AAT	11	N	0	
RT	201	34693	ACA	7	T	0	

1. CodFreq files contain tables with 7 columns containing the frequency of each codon at each position.
2. In this example the different codons present at positions 200 and 201 are shown.
3. Overall, there are approximately 34,000 reads encompassing these two positions.
4. Most of the codons are present at very low levels and represent background noise caused by PCR or machine errors.
5. <>The mutation T200A is present in about 21% of reads, well above the level of background noise.
6. <>Nearly all of the remaining codons shown on this slide, with the possible exception of the silent mutation present in 2% of reads at position 201, likely represent background noise.
7. There are three main advantages of working with codon frequency files.
8. <>First, they are much smaller than FASTQ files allowing drug resistance interpretations to be generated rapidly.
9. <>Second, the distribution in the frequency of non-consensus codons provides an important measure of quality control.
10. Specifically, it allows the proportion of positions with mixed nucleotides to be determined at each mutation detection threshold.
11. <>The one disadvantage of codon frequency files is that they do not indicate

whether sub-consensus mutations at different positions occur on the same reads.
12. However, this is a minor disadvantage because it is not possible to be sure of linkage between mutations that are not close to one another unless single genome amplification was performed prior to sequencing.
13. In addition, linkage of mutations is currently not considered during genotypic resistance interpretation.
14. <>Finally, CodFreq files have two advantages over the established variant call format or VCF files because they are interpretable without a reference sequence and can be used independently from an accompanying SAM or BAM file.

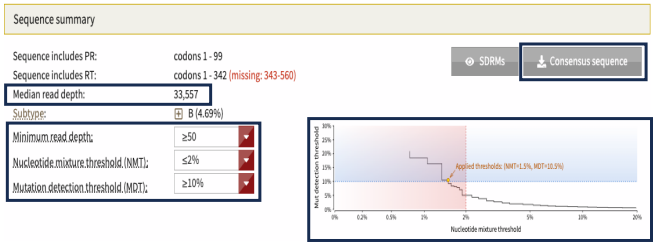
Unique aspects of the NGS output

View read coverage.

Adjust /re-adjust minimum read depth and mutation detection threshold.

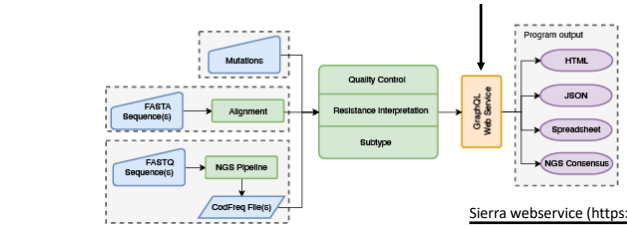
Download consensus FASTA sequence.

View proportions of sub-consensus variants.



1. This slide shows those aspects of the output that are unique to the analysis of NGS data.
2. <>The median read depth is shown and those parts of the sequence that have fewer reads than the minimum read depth are also indicated (although I am not showing those “warnings” on this slide).
3. <>After the interpretation is obtained, it is possible to reset the minimum read depth and two adjust the nucleotide mixture and amino acid mutation detection threshold.
4. <>Changes to these parameters will then be immediately reflected in the consensus sequence and in the resistance interpretation.
5. <>The output also indicates the read coverage at each of the positions containing a mutation and the proportion of reads for each mutation.
6. <>The figure on the right of the parameters shows how the nucleotide mixture threshold can override or suggest an appropriate mutation detection threshold.

Sierra Webservice



CodFreq pipeline (https://github.com/hivdb/codfreq)

- Open source on GitHub.
- Docker image is provided.

Sierra webservice (https://github.com/hivdb/sierra)

- Labs can access the GraphQL web service API to control how the output is formatted.
- Write their own client or use SierraPy.
- Open source on GitHub.
- Docker image is provided.

•1. This slide shows an overall schematic of the drug resistance interpretation programs.
•2. Users can submit a list of mutations, one or more FASTA sequences, or one or more CodFreq files.
•3. <>CodFreq files can be generated from FASTQ files using a program on our web site or using opensource code that we have developed.
•4. <>All of the output is delivered using a GraphQL webservice.
•5. <>Users who want to control the process of interacting with the drug resistance interpretation program and of formatting their output, they can access the GraphQL web service through its API.
•6. They can write their own client or use a client we developed called SierraPy.
•7. They can also run the entire program locally using the Sierra Docker image.

HIVDB Genotypic Drug Resistance Interpretation Program

For questions and suggestions:

hivdbteam@lists.Stanford.edu