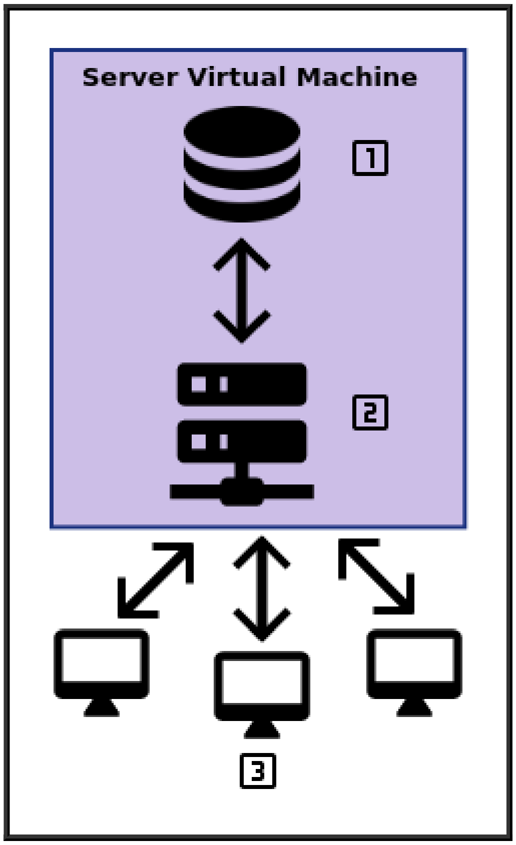
MaSTR version 1.4

Software Architecture



The software is structured into three parts, but the database (1) and api server (2) are both contained within the Server virtual machine image

1. The Database
    - Used to manage the job queue, store analysis results, and store important information (Panels, Models, Protocol Sets, etc.)
2. The API Server

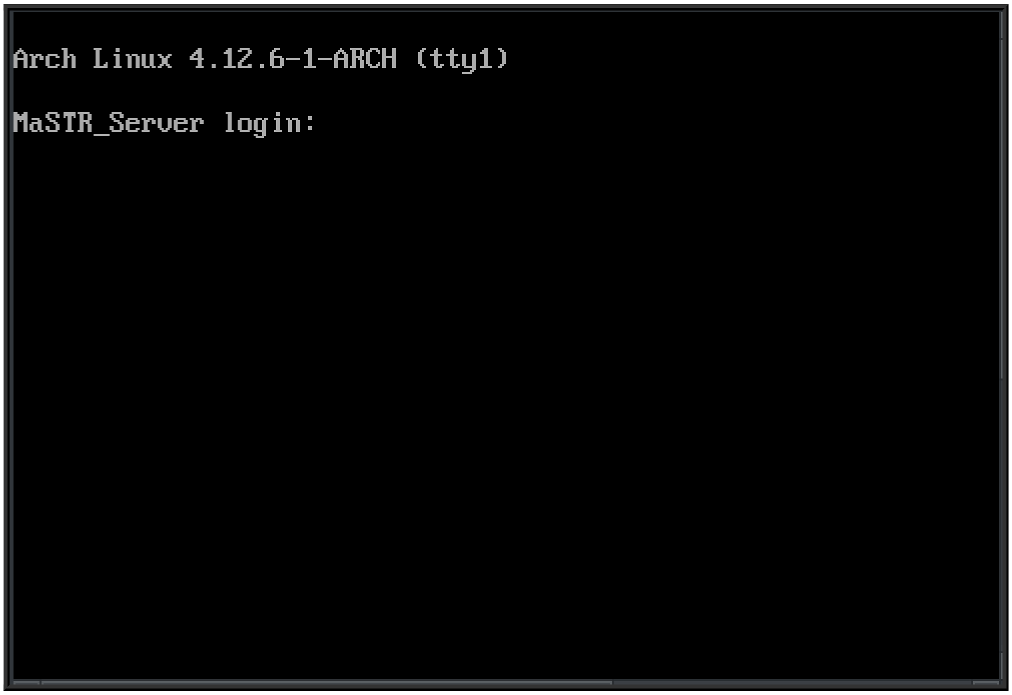
- Runs the analysis and communicates with the client to provide results

3. The Client
    - One or more clients may connect to the server to submit jobs and review results

Starting the Server

The server must be running before clients can connect and run an analysis.

The provided *.ova* file can be imported into VMWare or Virtualbox. The amount of RAM provided to the machine should be increased if possible. After launching the server VM a login console will be seen, but it is not necessary to login- the server will be accessible by clients within a few seconds.



```
Arch Linux 4.12.6-1-ARCH (tty1)

MaSTR_Server login:
```

The server virtual machine shows a login screen after booting up, but there is no need to login.

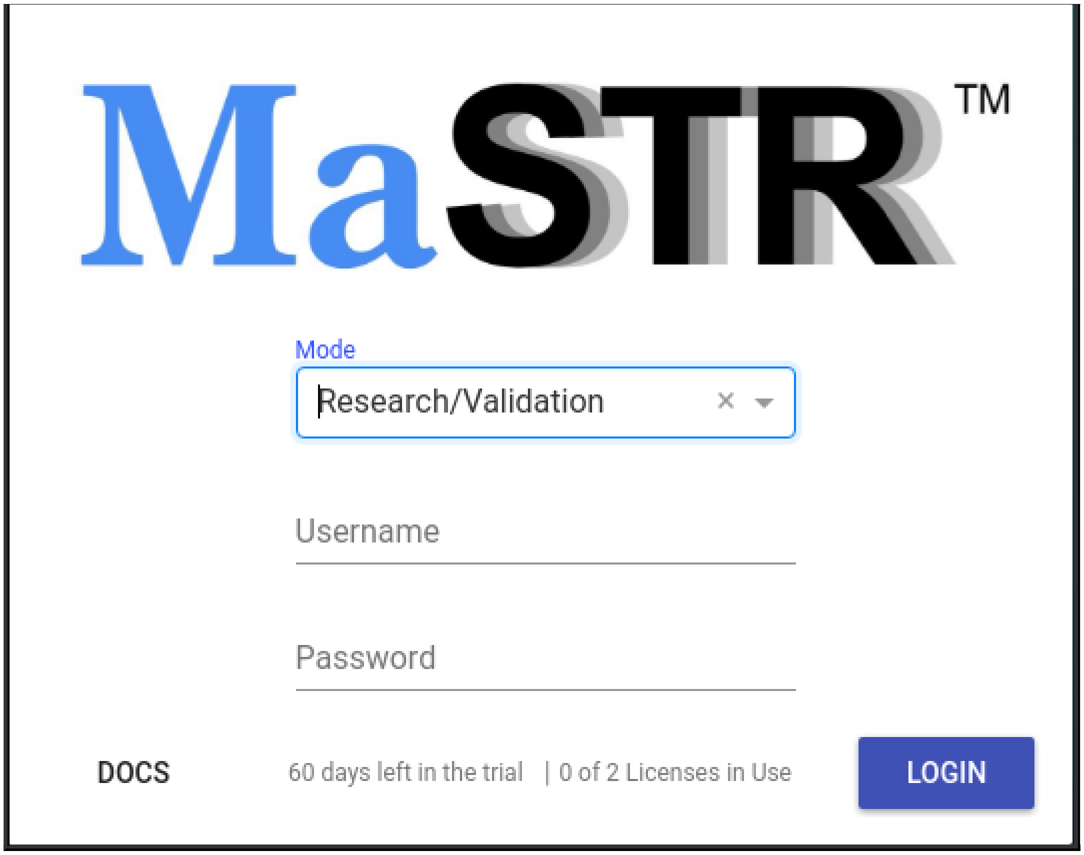
If it is necessary, settings in the virtualization program can be adjusted to change the port forwarding settings. By default the 8000 port on the server is forwarded to the 8000 port on the actual computer. This is the only port that needs to be exposed.

Using the Client

The client is accessible through a webbrowser (Google's Chrome™ browser is recommended). The server can be accessed on the computer running the virtual machine at port 8000 (or whichever port is exposed from the virtual machine). If the server is running on the same computer as the client, the location should be `https://localhost:8000` . It is necessary to allow an exception in order to trust the server's certificate which is used to encrypt traffic between the client and the server. When a user is logged in, a session is created.

The best way to learn how to use the client is to login in 'demo' mode (no username/password needed) and take the guided tour through the software.

Logging In



Mode: Research/Validation

60 days left in the trial | 0 of 2 Licenses in Use

The login dialog

When the client is opened the dialog that is shown should be used to enter credentials.

There are several accounts set up by default (**username**, *password*):

1. **root**, *changeThisPassword*
   - The root account is special. While it has the ability to access any APIs and perform user management, it isn't visible in user management and therefore cannot have its permissions or password modified by other users and cannot be deleted.
   - The root account is the only one with the ability to backup and restore the database.
2. **admin**, *admin*
   - Is a *User Administrator*, allowing login to the admin interface.
   - *manager* on the work and research databases

3. **researcher**, *researcher*
    - *researcher* on the work and research databases
4. **analyst**, *analyst*
    - *analyst* on the work and research databases
5. **reviewer**, *reviewer*
    - No special permissions on the work and research databases

> ⚠ **Warning**
>
> Before putting the software into production it is important to change the **root** password and either delete the other accounts or change their passwords.

> ✎ **Note**
>
> Logins expire after 8 hours. When a login is expired the user will be redirected to the login page. This will not affect any jobs already queued or running on the server.

Running an Analysis

See Running an Analysis

Models

Models specify how the software calculates expected peak heights and how the MCMC process is run.

Input Format

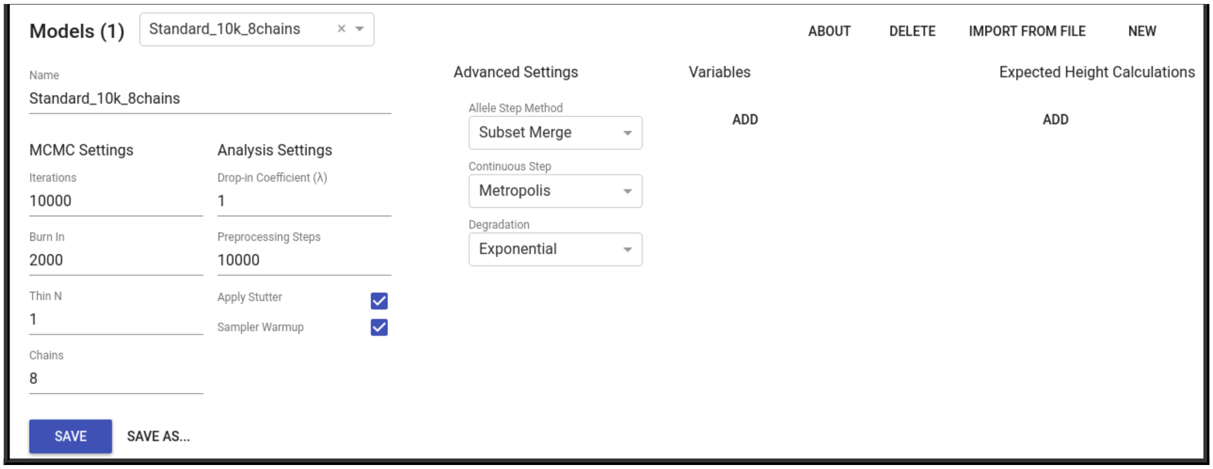
Models may be imported in json format, but are usually created within the software directly.

```
{
  "name" : "Marker-Contrib Adjustment",
  "mcmc_options" : {
    "burn_in" : 2000,
    "iterations" : 10000,
    "thin_n" : 2,
    "chains": 10
  },
  "analysis_options" : {
    "drop_in_coefficient": 1.0,
    "stutter":true,
    "sampler_warmup": true,
    "preprocessing_steps": 10000
  },
  "allele_step_method" : "Subset Merge",
  "continuous_step" : "Metropolis",
  "degradation": "Linear",
  "variables" : [
    {
      "params" : {
        "value" : 1
      },
      "description" : "Set a value of 1 for each marker to shape a
child variable as marker-contributor",
      "shape" : "marker",
      "dist" : "Fixed",
      "name" : "Marker_Fixed_1"
    },
    {
      "params" : {
        "mean" : "Marker_Fixed_1",
        "sd" : 0.2,
        "lower" : 0.5,
        "upper" : 1.5
      },
      "description" : "",
      "shape" : "contributor",
      "dist" : "Bounded_Normal",
      "name" : "Contrib-Marker"
    }
  ],
  "expected_height_calc" : [
      [
        "Multiply",
        "Contrib-Marker"
      ]
  ]
}
```

The model name must be unique in the database.

Models in MaSTR

Models can be imported, viewed, modified, and deleted in the "Models" view.



Standard_10k_8chains

Name: Standard_10k_8chains

MCMC Settings

- Iterations: 10000
- Burn In: 2000
- Thin N: 1
- Chains: 8

Analysis Settings

- Drop-in Coefficient (λ): 1
- Preprocessing Steps: 10000
- Apply Stutter ☑
- Sampler Warmup ☑

Advanced Settings

- Allele Step Method: Subset Merge
- Continuous Step: Metropolis
- Degradation: Exponential

Variables

Expected Height Calculations

MCMC Settings

These settings affect the MCMC process.

Analysis Settings

The "Drop-in Coefficient" setting adjusts the probability of drop-in.

The *Apply Stutter* setting may enable/disable stutter as specified by the protocol set.

It is possible to perform Sampler Warmup for a specified number of *preprocessing steps*.

Advanced Settings

The *Allele Step Method* and *Continuous Step Method* settings relate to the details of the MCMC sampling process.

Degradation can be set to "No Adjustment", "Linear", or "Exponential".

Model Variables

Model Variables allow for much more complex probabilistic models to be specified in analyzing the data. They may be created with the *ADD* button under "Variables" before being inserted into the "Expected Height Calculation".

About

The *ABOUT* button provides any notes about the model in addition to information on its creation and modification

Panels

Panels contain information about the kit used to generate any signal data and/or profiles.

Input Format

Standard XML-formatted panel files (like those used in GeneMarker) can be imported into MaSTR. A partial example:

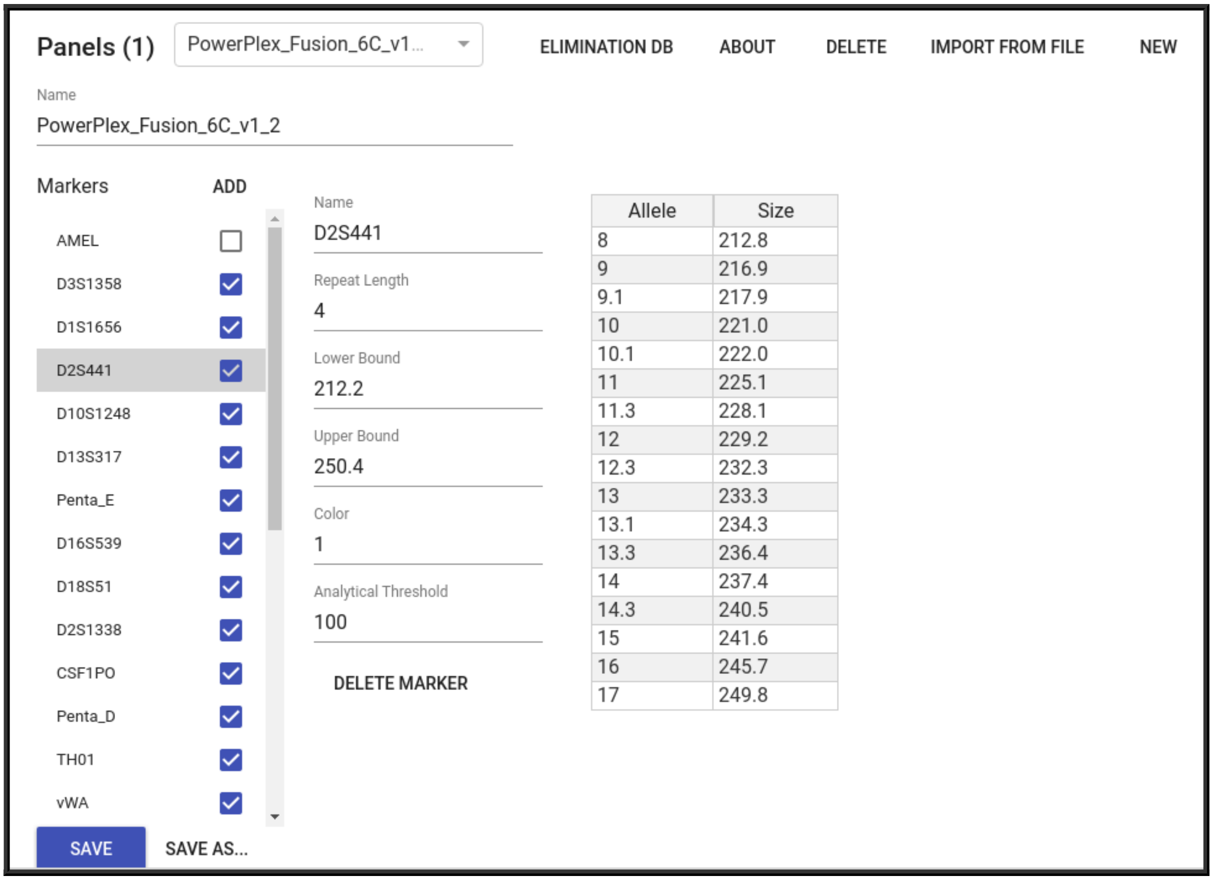
```
<?xml version="1.0"?>
<Panel>
    <Vendor>GeneMarker</Vendor>
    <FileVersion>01.3a</FileVersion>
    <PanelName>PowerPlex_16</PanelName>
    <n_Ploidy>2</n_Ploidy>
    <Chemistry></Chemistry>
    <Lot_Number></Lot_Number>
    <Fragment Enabled = '0'> </Fragment>
    <Loci>
        <Locus>
        <MarkerTitle>D3S1358</MarkerTitle>
        <DyeIndex>1</DyeIndex>
        <n_NucleotideRepeats>4</n_NucleotideRepeats>
        <Enabled>1</Enabled>
        <LowerBoundary>106.5</LowerBoundary>
        <UpperBoundary>149.4</UpperBoundary>
        <MarkerComments></MarkerComments>
        <LocusFilter    MinIntensity = '50' HomoInconclusive = '200'
MinHeteroIntensity = '50'   HeteroInconclusive = '100'  LocalPercent
= '20' DecimalLocalPercent = '0'   StutterPer_N_L4 = '13'
DecimalStutterPer_N_L4 = '0'    StutterPer_N_L8 = '1'
DecimalStutterPer_N_L8 = '0'    StutterPer_N_R4 = '1'
DecimalStutterPer_N_R4 = '0'    HetroImbalancePer = '60'
```

```
DecimalHetroImbalancePer = '0'  ></LocusFilter>
        <Allele     Label = '11'    DefSize = '106.20'  Size =
'107.08' Left_Binning = '0.50'   Right_Binning = '0.50'  Control =
'0'   Distance = '0.00'   Chrom_Pos = ''   Recom_Freq = '-1.00'
Height = '90'   Comments = ''   ></Allele>
        <Allele     Label = '12'    DefSize = '110.42'  Size =
'111.40' Left_Binning = '0.50'   Right_Binning = '0.50'  Control =
'1'   Distance = '0.00'   Chrom_Pos = ''   Recom_Freq = '-1.00'
Height = '126'  Comments = ''   ></Allele>
```

The panel name and version must be unique in the database.

Panels in MaSTR

Panels can be imported, viewed, modified, and deleted on the "Panels" tab in the "System" view.



Name
PowerPlex_Fusion_6C_v1_2

- AMEL □
- D3S1358 ☑
- D1S1656 ☑
- D2S441 ☑
- D10S1248 ☑
- D13S317 ☑
- Penta_E ☑
- D16S539 ☑
- D18S51 ☑
- D2S1338 ☑
- CSF1PO ☑
- Penta_D ☑
- TH01 ☑
- vWA ☑

Name
D2S441

Repeat Length
4

Lower Bound
212.2

Upper Bound
250.4

Color
1

Analytical Threshold
100

Allele	Size
8	212.8
9	216.9
9.1	217.9
10	221.0
10.1	222.0
11	225.1
11.3	228.1
12	229.2
12.3	232.3
13	233.3
13.1	234.3
13.3	236.4
14	237.4
14.3	240.5
15	241.6
16	245.7
17	249.8

The important panel information includes:

- Which markers are present and enabled
- Which alleles are possible in those markers
- The analytical threshold to use for each marker

A panel can be selected from the drop-down menu. Panels must have a unique combination of name and version.

Markers

The left side lists the markers found in the panel, with checkboxes to enable or disable use of those markers.

> ⚠ **Warning**
>
> Amelogenin is disabled by default, regardless of whether or not the panel has this marker enabled. Specifically, the software checks for 'amel' in the name of each marker.

> ⚠ **Warning**
>
> Spaces in marker names are replaced with underscores. For example, "Penta D" becomes "Penta_D".

Elimination Profiles

The *ELIMINATION DB* button allows for elimination profiles to be uploaded, modified, or deleted. See Elimination Database

About

The *ABOUT* button provides any notes about the panel in addition to information on its creation and modification

Frequencies

Frequencies contain population frequency for alleles in different markers and are organized according to publication.

Input Format

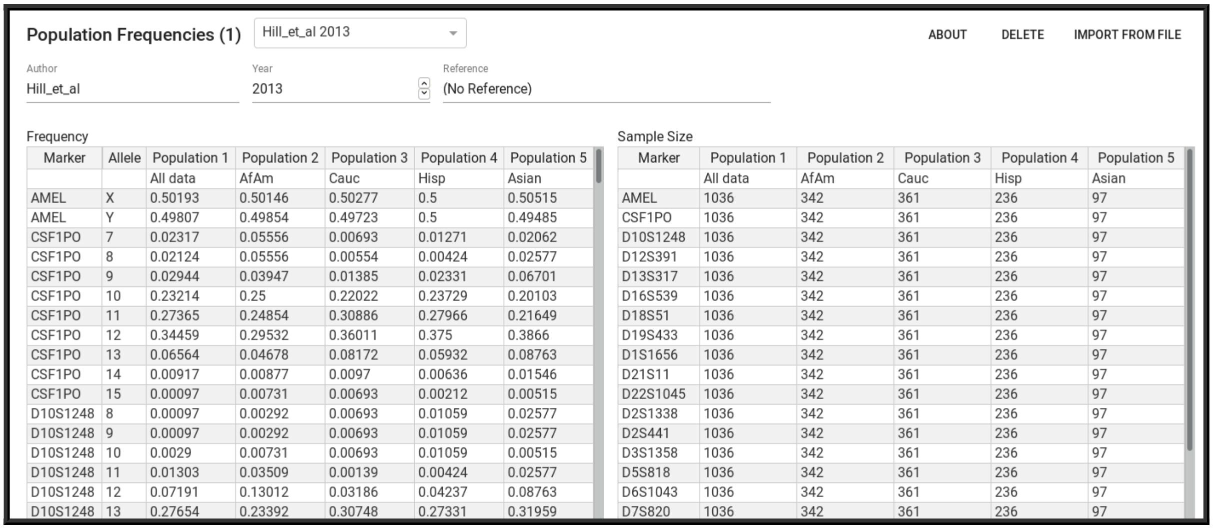
When importing a frequency file, the software expects a file with 3 header lines with Author, Year, and Reference information followed by a tab-delimited listing of marker name, allele name, and population frequency. The first allele listed for each marker should be 'N' and instead of a frequency the file should specify the sample number.

The author name and year must be unique in the database.

```
#Author Hill_et_al
#Year   2013
#Reference  (No Reference)
Marker  Allele  All data    AfAm    Cauc    Hisp    Asian
CSF1PO  N   1036    342 361 236 97
CSF1PO  7   0.02317 0.05556 0.00693 0.01271 0.02062
CSF1PO  8   0.02124 0.05556 0.00554 0.00424 0.02577
CSF1PO  9   0.02944 0.03947 0.01385 0.02331 0.06701
CSF1PO  10  0.23214 0.25000 0.22022 0.23729 0.20103
CSF1PO  11  0.27365 0.24854 0.30886 0.27966 0.21649
CSF1PO  12  0.34459 0.29532 0.36011 0.37500 0.38660
CSF1PO  13  0.06564 0.04678 0.08172 0.05932 0.08763
CSF1PO  14  0.00917 0.00877 0.00970 0.00636 0.01546
CSF1PO  15  0.00097 0.00731 0.00693 0.00212 0.00515
D10S1248    N   1036    342 361 236 97
D10S1248    8   0.00097 0.00292 0.00693 0.01059 0.02577
```

Frequencies in MaSTR

Panels can be imported, viewed, modified, and deleted on the "Frequencies" tab in the "System" view.



Author: Hill_et_al

Year: 2013

Reference: (No Reference)

Frequency

Marker	Allele	Population 1	Population 2	Population 3	Population 4	Population 5
		All data	AfAm	Cauc	Hisp	Asian
AMEL	X	0.50193	0.50146	0.50277	0.5	0.50515
AMEL	Y	0.49807	0.49854	0.49723	0.5	0.49485
CSF1PO	7	0.02317	0.05556	0.00693	0.01271	0.02062
CSF1PO	8	0.02124	0.05556	0.00554	0.00424	0.02577
CSF1PO	9	0.02944	0.03947	0.01385	0.02331	0.06701
CSF1PO	10	0.23214	0.25	0.22022	0.23729	0.20103
CSF1PO	11	0.27365	0.24854	0.30886	0.27966	0.21649
CSF1PO	12	0.34459	0.29532	0.36011	0.375	0.3866
CSF1PO	13	0.06564	0.04678	0.08172	0.05932	0.08763
CSF1PO	14	0.00917	0.00877	0.0097	0.00636	0.01546
CSF1PO	15	0.00097	0.00731	0.00693	0.00212	0.00515
D10S1248	8	0.00097	0.00292	0.00693	0.01059	0.02577
D10S1248	9	0.00097	0.00292	0.00693	0.01059	0.02577
D10S1248	10	0.0029	0.00731	0.00693	0.01059	0.00515
D10S1248	11	0.01303	0.03509	0.00139	0.00424	0.02577
D10S1248	12	0.07191	0.13012	0.03186	0.04237	0.08763
D10S1248	13	0.27654	0.23392	0.30748	0.27331	0.31959

Sample Size

Marker	Population 1	Population 2	Population 3	Population 4	Population 5
	All data	AfAm	Cauc	Hisp	Asian
AMEL	1036	342	361	236	97
CSF1PO	1036	342	361	236	97
D10S1248	1036	342	361	236	97
D12S391	1036	342	361	236	97
D13S317	1036	342	361	236	97
D16S539	1036	342	361	236	97
D18S51	1036	342	361	236	97
D19S433	1036	342	361	236	97
D1S1656	1036	342	361	236	97
D21S11	1036	342	361	236	97
D22S1045	1036	342	361	236	97
D2S1338	1036	342	361	236	97
D2S441	1036	342	361	236	97
D3S1358	1036	342	361	236	97
D5S818	1036	342	361	236	97
D6S1043	1036	342	361	236	97
D7S820	1036	342	361	236	97

About

The *ABOUT* button provides any notes about the frequency data in addition to information on its creation and modification

Protocol Sets

Protocol Sets serve two purposes:

1. Store information related to the SOP used for analysis (instrument, injection time, etc)
2. Use single-source data to measure both stutter and variance

Input Format

Protocol sets may be imported in json format, but are usually created within the software directly. The protocol set name must be unique in the database.

A protocol set requires single-source data and the corresponding genotypes in order to measure the variation and calculate stutter values.

Data:

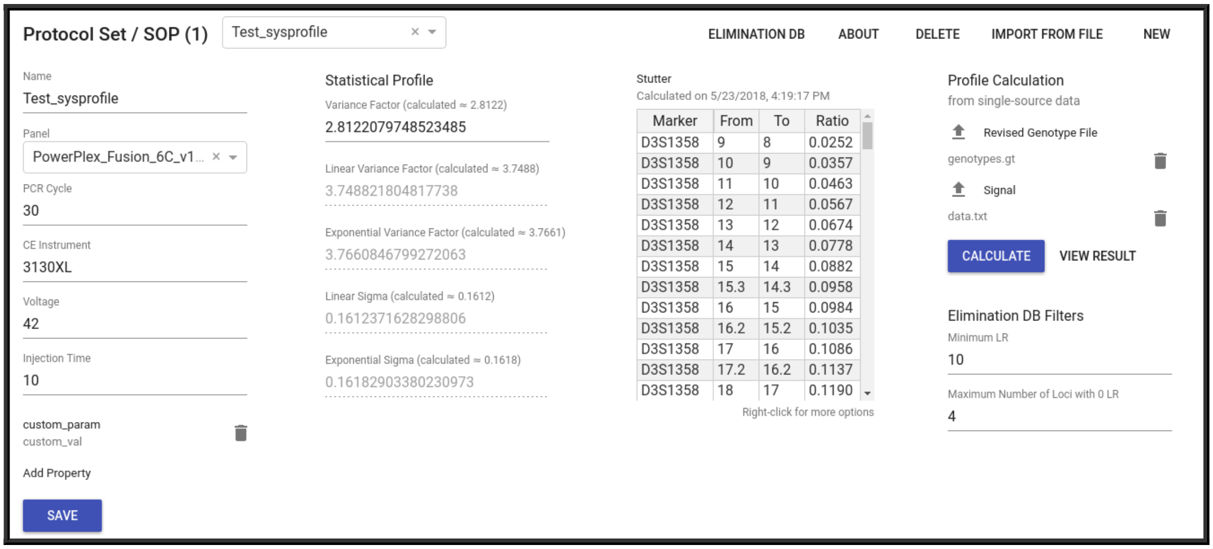
```
Sample Name Marker  Allele#1    Allele#2    Allele#3    Allele#4
Allele#5    Height#1    Height#2    Height#3    Height#4    Height#5
Sample123   D3S1358 14.3    15  15.3    16  16.2    14  33  43  778
11
Sample123   D1S1656 11  12  13  14  15              15  277 10  290
296
Sample123   D2S441  11  12  13                      24  213 110
Sample123   D10S1248    12  13  14  15              180 305 17  188
```

Genotypes:

```
Sample Name Marker  Allele#1    Allele#2    Height#1    Height#2
Sample123   AMEL    X   X   677 677
Sample123   D3S1358 16  16  778 778
Sample123   D1S1656 12  15  277 296
Sample123   D2S441  12  14  213 264
Sample123   D10S1248    13  15  305 188
Sample123   D13S317 10  11  250 180
```

Protocol Sets in MaSTR

Protocol sets can be imported, viewed, modified, and deleted on the "Protocol Sets" tab in the "System" view.



Name
Test_sysprofile

Panel
PowerPlex_Fusion_6C_v1...

PCR Cycle
30

CE Instrument
3130XL

Voltage
42

Injection Time
10

custom_param
custom_val

Statistical Profile

Variance Factor (calculated ≈ 2.8122)
2.8122079748523485

Linear Variance Factor (calculated ≈ 3.7488)
3.748821804817738

Exponential Variance Factor (calculated ≈ 3.7661)
3.7660846799272063

Linear Sigma (calculated ≈ 0.1612)
0.1612371628298806

Exponential Sigma (calculated ≈ 0.1618)
0.16182903380230973

Stutter
Calculated on 5/23/2018, 4:19:17 PM

Marker	From	To	Ratio
D3S1358	9	8	0.0252
D3S1358	10	9	0.0357
D3S1358	11	10	0.0463
D3S1358	12	11	0.0567
D3S1358	13	12	0.0674
D3S1358	14	13	0.0778
D3S1358	15	14	0.0882
D3S1358	15.3	14.3	0.0958
D3S1358	16	15	0.0984
D3S1358	16.2	15.2	0.1035
D3S1358	17	16	0.1086
D3S1358	17.2	16.2	0.1137
D3S1358	18	17	0.1190

Profile Calculation
from single-source data

Revised Genotype File
genotypes.gt

Signal
data.txt

Elimination DB Filters

Minimum LR
10

Maximum Number of Loci with 0 LR
4

Each protocol set is associated with a Panel that has previously been added to the server.

SOP Info

- Panel - each SOP is associated with a single panel.
- PCR Cycle
- CE Instrument - it may be beneficial to associate all data run on a single instrument in order to reduce variability.

- Voltage
- Injection Time
- Custom Parameters - custom parameter names and values can be added to track any additional information

Profile Calculation

After loading single-source data and matching profiles, the *CALCULATE* button will determine the variance, stutter, and degradation values.

Elimination DB Settings

Profiles in the elimination database for the selected panel are flagged when they exceed the specified likelihood ratio and have fewer than the maximum number of 0 LR markers.

About

The *ABOUT* button provides any notes about the protocol set in addition to information on its creation and modification

Profile and Signal Files

Both of these files types are tab-delimited files with a single header line and one row per marker.

Signal Files

Signal files contain the allele and peak height information for a sample. They should include at least one non-OB allele for each marker, and the corresponding heights for each allele.

```
Marker  Allele#1    Allele#2    Allele#3    Allele#4    Height#1
Height#2    Height#3    Height#4
AMEL    X   Y           2544    510
D3S1358 15  16  17  18  189 881 665 356
D1S1656 11  14  16      945 218 320
D2S441  10  11  13  14  237 531 300 282
```

Profile Files

Profile files contain the allele information for a potential/known contributor and optionally also include the height.

```
Marker  Allele#1    Allele#2
AMEL    X   X
D3S1358 16  17
D1S1656 13  16
D2S441  11  11
```

Sample Names

Either file type may also include a "Sample Name" in each line. Files with multiple profiles or samples must include sample names. This type of file may be used in the elimination database and is always used for the system profile calculation.

```
Sample Name Marker  Allele#1    Allele#2    Height#1    Height#2
Profile_GGC AMEL    X   X   677 677
Profile_GGC D3S1358 16  16  778 778
Profile_GGC D1S1656 12  15  277 296
Profile_GGC D2S441  12  14  213 264
```

Note

When a sample name is not included in a file, the file's name is used to name that sample or signal data.

Note

When a profile or signal file loaded into a job contains more than one Sample Name, only data from the first sample is loaded.

Profile Types

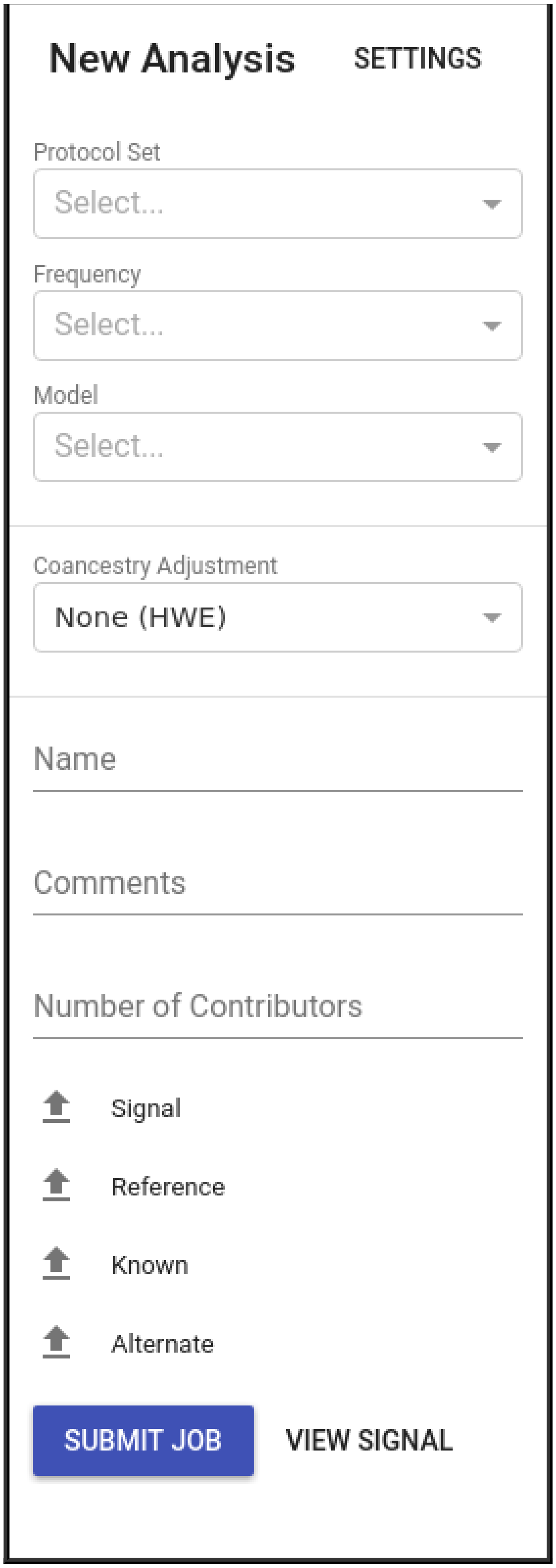
- Reference
  - The profile being tested in the job
- Known
  - A profile that is assumed to be in the mixture.
- Alternate
  - Alternate profiles are calculated as if they were the reference file. They are most useful when testing LRs for known mixtures.
- Elimination

- Elimination profiles are stored with an associated panel. They are always used to calculate LRs against the mixture, and are reported if these results exceed specified thresholds.

Running an analysis

Submitting a new job

The 'NEW' button opens a dialog used to submit a new job to the server.



Coancestry Adjustment

None (HWE)

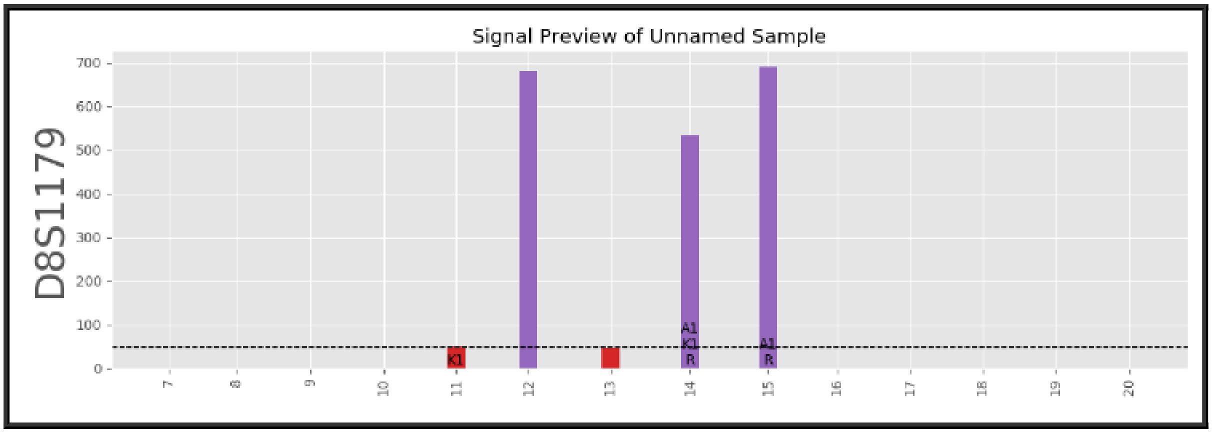
Used to create and submit a new job

1. Select a Protocol Set

2. Select a Frequency
3. Select a Model
4. Select a method and value for Coancestry Adjustment
5. Enter a `Name` and any `Comments`
6. Set the `Number of Contributors`
7. Load the data to be analyzed
    - A signal file containing the actual mixture data
    - A reference profile to be tested
    - Any known profiles that are assumed to be present in the mixture
    - Any alternate profiles that will be tested as a reference using the same MCMC results

The *SETTINGS* button in the upper-right of this dialog allows changes to the MCMC and analysis settings without changing the original model. This is available only in 'demo' and 'research' modes.

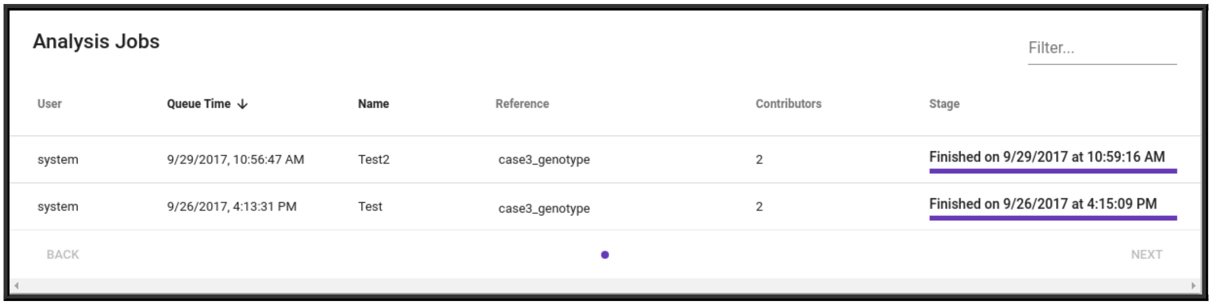
The *VIEW SIGNAL* button uses the signal and any genotypes (reference, known, and/or alternate) and graphs the peak heights. Each peak is labeled with the genotypes that have that allele (K# = Known, A# = Alternate, and R = Reference). The analytical threshold is shown and any peaks underneath it are colored red.



A plot of D8S1179 signal heights with annotated peaks for the reference (14, 15), a known (11, 14) and an alternate (14, 15). One of the known *peaks (11) and another peak (13) are below the analytical threshold- those peaks would not be included in the analysis.*

Job List

After login, the main page of the client shows a list of jobs (finished, running, or queued) and some basic information about each one.



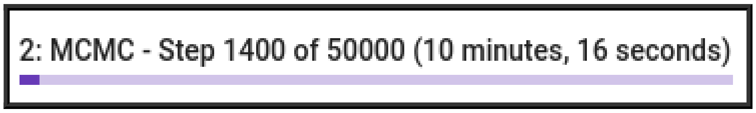
User	Queue Time ↓	Name	Reference	Contributors	Stage
system	9/29/2017, 10:56:47 AM	Test2	case3_genotype	2	Finished on 9/29/2017 at 10:59:16 AM
system	9/26/2017, 4:13:31 PM	Test	case3_genotype	2	Finished on 9/26/2017 at 4:15:09 PM

List of jobs in the ANALYSIS tab

When running, jobs will report the current stage of the analysis and each stage has its own progress bar. There are four stages:

1. Setup
2. MCMC
3. Calculate LR
4. Collect and Save Results

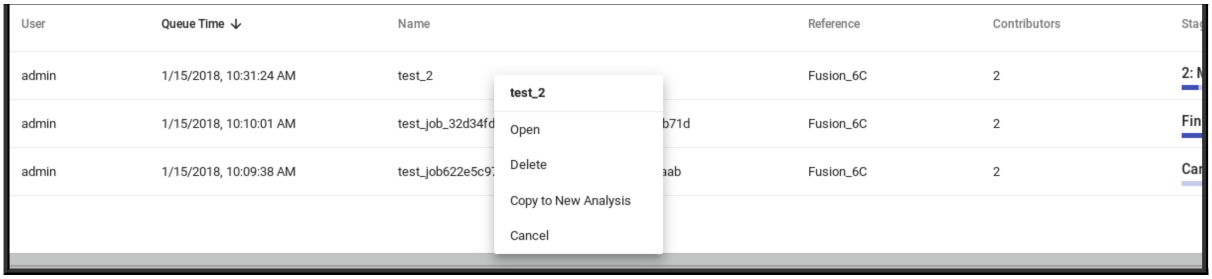
During the MCMC stage the step number and an estimated remaining time is shown.



2: MCMC - Step 1400 of 50000 (10 minutes, 16 seconds)

Status as seen while a project is in the MCMC stage of analysis

Clicking on a job in the list will open a results dialog. Right-clicking on a job will list several options, depending on the current job status and user permissions:



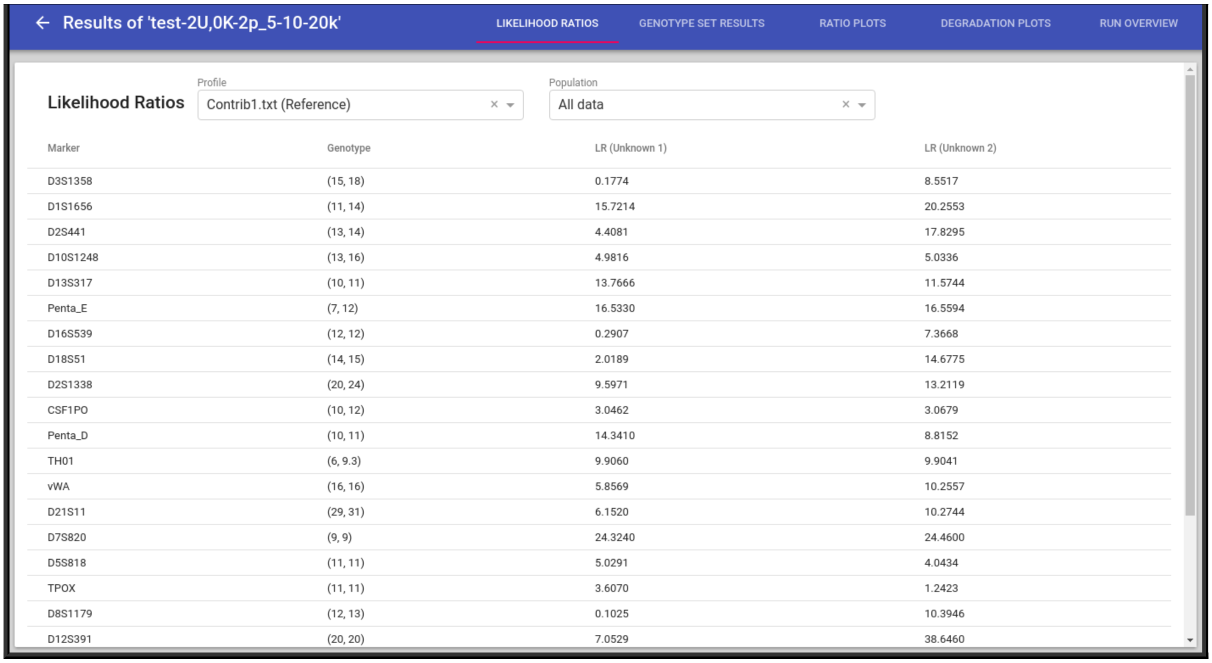
User	Queue Time ↓	Name	Reference	Contributors	Stag
admin	1/15/2018, 10:31:24 AM	test_2	Fusion_6C	2	2: M
admin	1/15/2018, 10:10:01 AM	test_job_32d34fd b71d	Fusion_6C	2	Fin
admin	1/15/2018, 10:09:38 AM	test_job622e5c97 aab	Fusion_6C	2	Car

Right-click menu of options available for a currently-running job

- *Open* - Open the results dialog
- *Delete* - If the user has the required permission, removes the job from the database.
- *Copy to New Analysis* - Opens a New Analysis dialog with the settings from the selected job. Useful for running multiple similar jobs.
- *Download Report* - Download a pdf report for the job's results.
- *Cancel* - Cancel a job that is queued or currently running. Users may only cancel jobs that they have created.

Results

Clicking on a job in the job list opens a results dialog. Here, the results can be used to copy the inputs and settings for creating a new run (*COPY TO NEW ANALYSIS*) or downloaded as a PDF report (*DOWNLOAD REPORT*). If the user has permission to delete jobs, a (*DELETE*) button is also present.



Likelihood Ratios

Profile: Contrib1.txt (Reference)

Population: All data

Marker	Genotype	LR (Unknown 1)	LR (Unknown 2)
D3S1358	(15, 18)	0.1774	8.5517
D1S1656	(11, 14)	15.7214	20.2553
D2S441	(13, 14)	4.4081	17.8295
D10S1248	(13, 16)	4.9816	5.0336
D13S317	(10, 11)	13.7666	11.5744
Penta_E	(7, 12)	16.5330	16.5594
D16S539	(12, 12)	0.2907	7.3668
D18S51	(14, 15)	2.0189	14.6775
D2S1338	(20, 24)	9.5971	13.2119
CSF1PO	(10, 12)	3.0462	3.0679
Penta_D	(10, 11)	14.3410	8.8152
TH01	(6, 9.3)	9.9060	9.9041
vWA	(16, 16)	5.8569	10.2557
D21S11	(29, 31)	6.1520	10.2744
D7S820	(9, 9)	24.3240	24.4600
D5S818	(11, 11)	5.0291	4.0434
TPOX	(11, 11)	3.6070	1.2423
D8S1179	(12, 13)	0.1025	10.3946
D12S391	(20, 20)	7.0529	38.6460

Results for a 2-person mixture

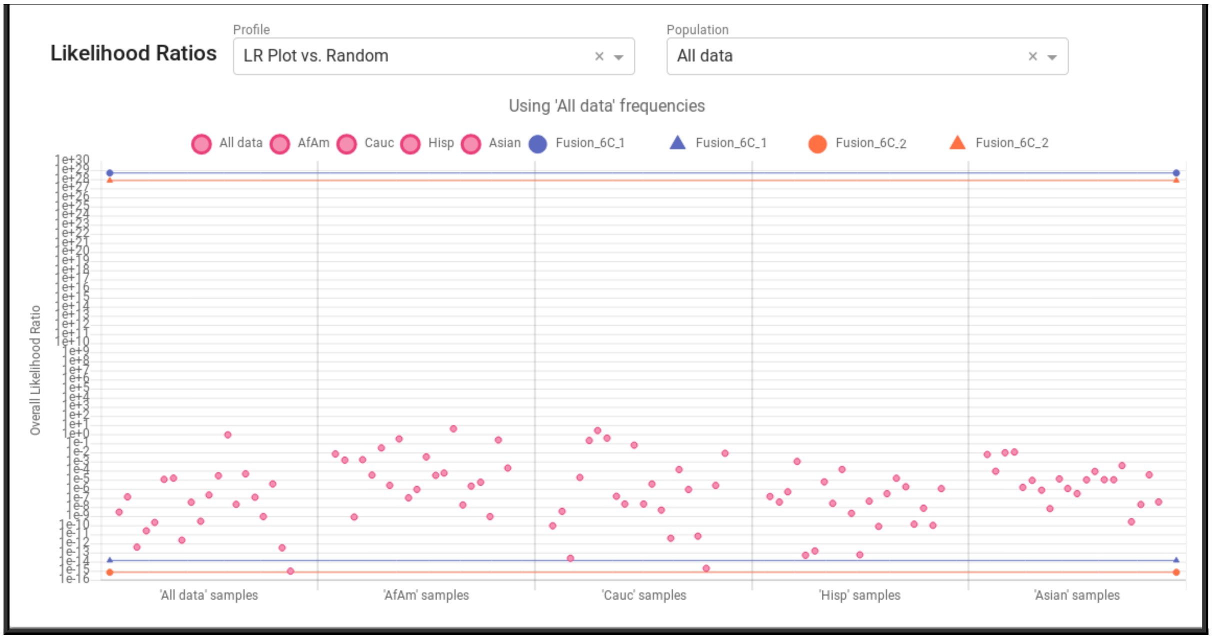
The results dialog includes several tabs:

Likelihood Ratios

This tab displays LR for each marker and the overall value. The LR can be calculated with any population from the frequency file using the "Population" drop-down selector. The other drop-down selector (*Profile*) allows the user to choose between:

- Results for the reference profile
- Results for any alternate profiles
- Results for any detected profiles from the elimination database
- A plot of overall LR results from randomly generated profiles

The overall likelihood ratios do not consider markers with an LR of 0 (which have red text).



Likelihood Ratios

Profile: LR Plot vs. Random

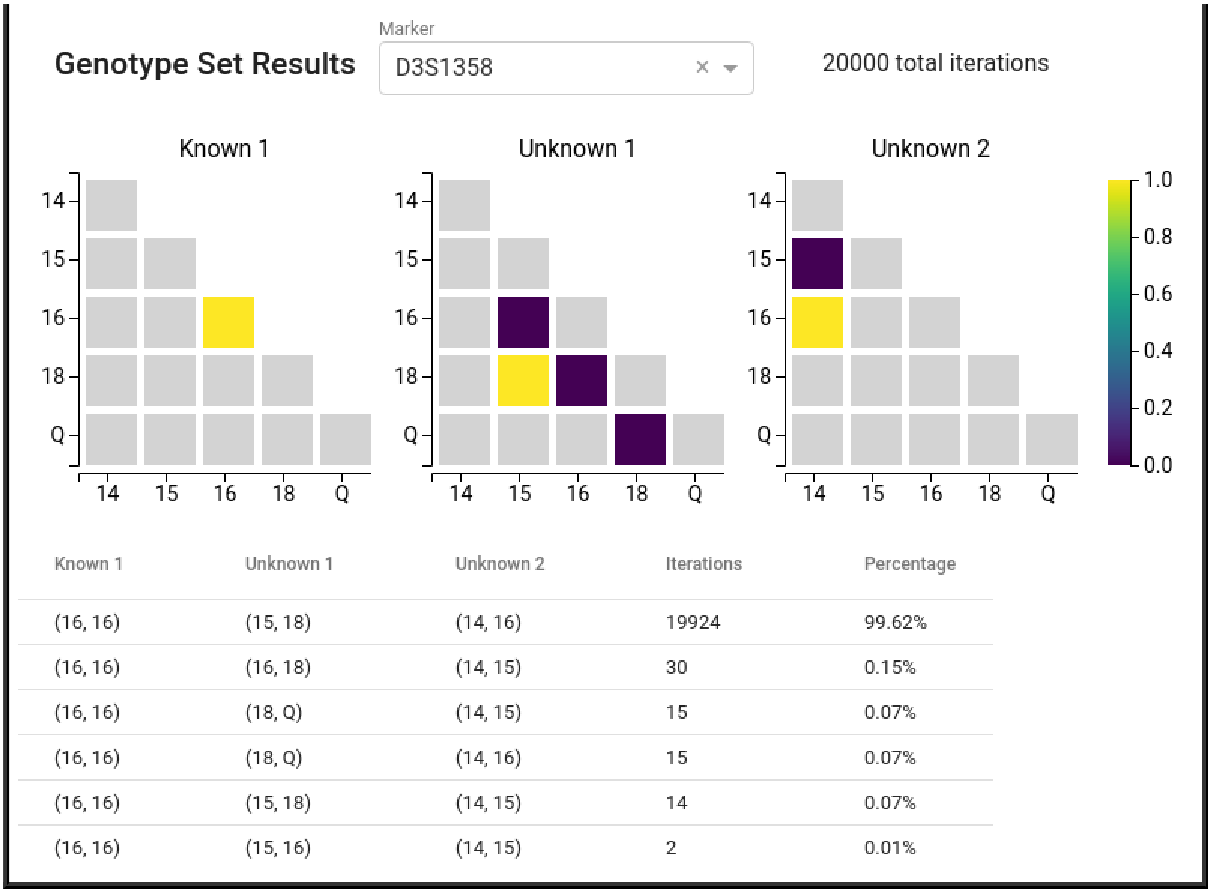
Population: All data

LR Plot vs Random

Selecting *LR Plot vs Random* on the *Profile* drop-down plots the maximum overall LR for random profiles from each population. These values are shown as dots, and the overall LR from any tested profiles as lines (with different colors for each profile and different shapes for each position in the genotype set). This provides some context for the overall LR values.

Genotype Set Results

This tab reports the relative probability of different genotype combinations according to how frequently they were accepted during the MCMC process.



Genotype Set Results

Marker

D3S1358

20000 total iterations

Known 1	Unknown 1	Unknown 2	Iterations	Percentage
(16, 16)	(15, 18)	(14, 16)	19924	99.62%
(16, 16)	(16, 18)	(14, 15)	30	0.15%
(16, 16)	(18, Q)	(14, 15)	15	0.07%
(16, 16)	(18, Q)	(14, 16)	15	0.07%
(16, 16)	(15, 18)	(14, 15)	14	0.07%
(16, 16)	(15, 16)	(14, 15)	2	0.01%

This tab shows the frequency of the sampled genotypes

Ratio Plots

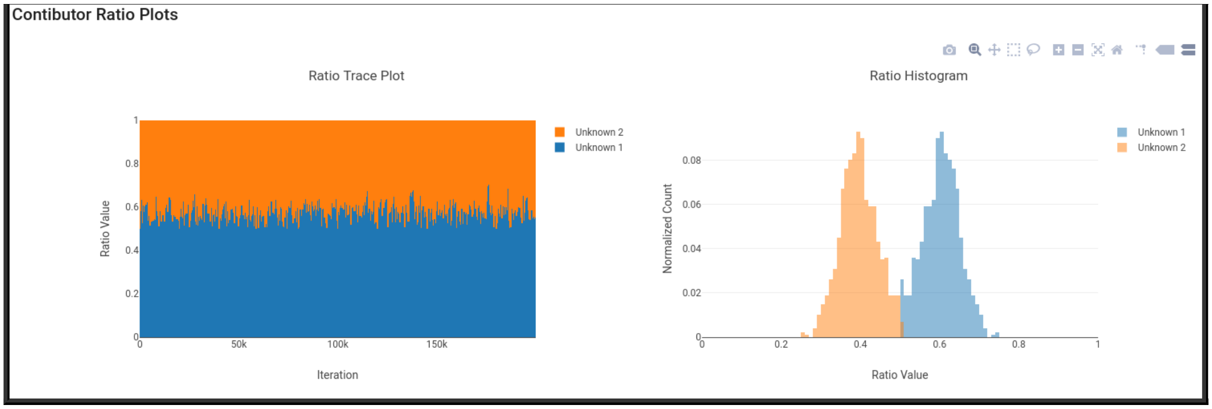
This tab includes plots for the sampled ratio values for each contributor- a plot of the values in each trace and a histogram of the values.

Note

The ratio, degradation, and model variable plots actually only display 1,000 datapoints evenly sampled from the trace. The original data is stored in the database, but this subset is returned to the client to save time.

Note

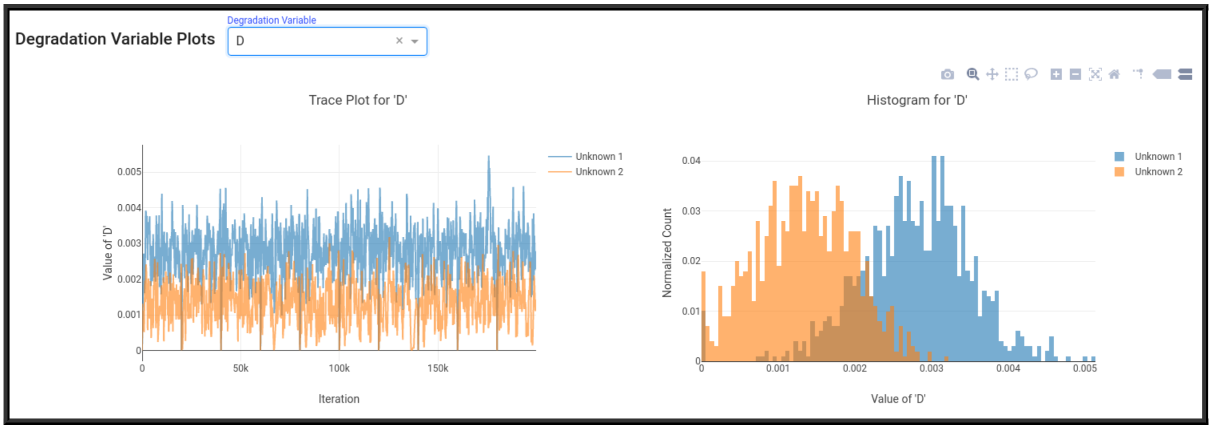
The ratio, degradation, and model variable plots are interactive- It's possible to click on labels in the plot legends to toggle their display.


Contibutor Ratio Plots

A Ratio plot showing the estimated ratios throughout the trace and a histogram

Degradation Plots

These plots show the values of degradation variables.


Degradation Variable Plots

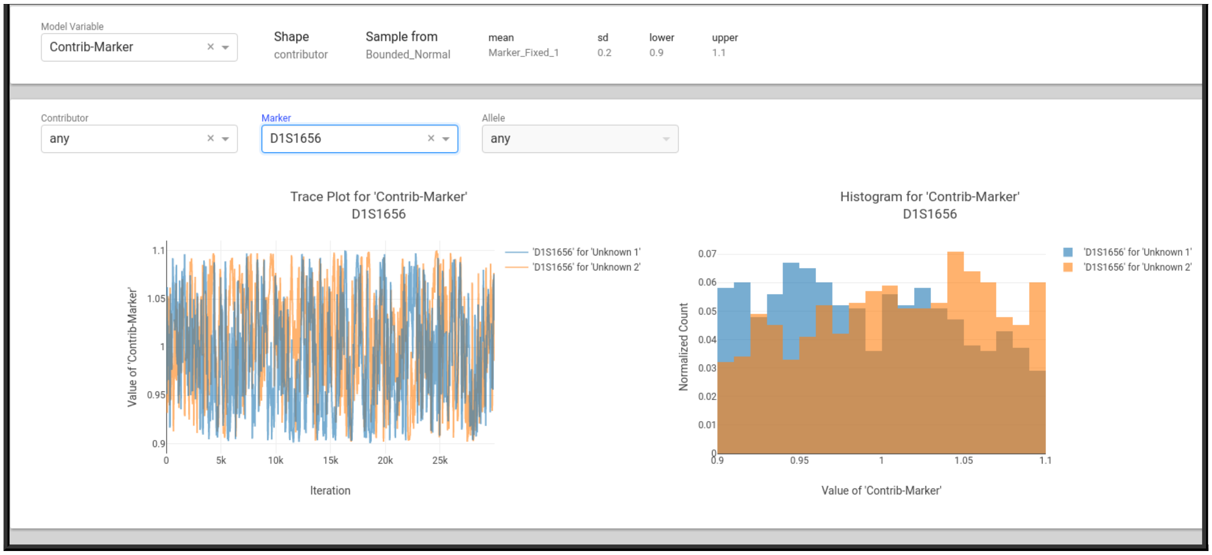
Degradation Variable: D

A plot of degradation variable 'D' showing the estimated values throughout the trace and a histogram

Model Variable Plots

This tab is only shown if the model contains model variables that are sampled as part of the MCMC process.

The top portion of this view allows model variables to be selected and displays relevant information. The bottom portion shows a trace plot and a histogram along with drop-down selectors to limit the plots to specific contributors, markers, and/or alleles.


Model Variable: Contrib-Marker

Shape: contributor

Sample from: Bounded_Normal

mean: Marker_Fixed_1

sd: 0.2

lower: 0.9

upper: 1.1

Contributor: any

Marker: D1S1656

Allele: any

Trace Plot for 'Contrib-Marker'
D1S1656
Value of 'Contrib-Marker'
1.1
1.05
1
0.95
0.9
0
5k
10k
15k
20k
25k
Iteration
'D1S1656' for 'Unknown 1'
'D1S1656' for 'Unknown 2'
Histogram for 'Contrib-Marker'
D1S1656
Normalized Count
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
0.9
0.95
1
1.05
1.1
Value of 'Contrib-Marker'
'D1S1656' for 'Unknown 1'
'D1S1656' for 'Unknown 2'

A plot of a variable that has a different value for each contributor and marker. It has been filtered to only show results for D1S1656

Run Overview

This tab contains some basic information about the run. It is possible to click the input data or settings to see all of the values that were used to create the run.

Backups

The admin interface includes the ability to download and restore full backups of the database. This can be accessed on the *BACKUP* tab when logged in to *admin* mode as the *root* user.

Note

Only the *root* user will see this tab. This is the only account with the ability to create or restore backups.

There are 3 buttons:

- BACKUP - Download a backup of the database
- VALIDATE - Check a backup file to ensure it is able to be restored
- RESTORE - Clear the current database, validate the new backup, and restore it

Permissions

Databases and Login Modes

Users may login in one of four modes:

- *Admin* - provides access to user management and other system functions
- *Demo* - provides unrestricted access to a temporary database and guides the user through the use of the software (no username and password needed).
- *Research/Validation* - a permanent database meant to be used for research and/or validation purposes.
- *Casework* - a permanent database meant to be used for casework-related work.

> **Note**
>
> The information stored in the database is not automatically backed up and is not (currently) able to be transferred between different versions of the software.

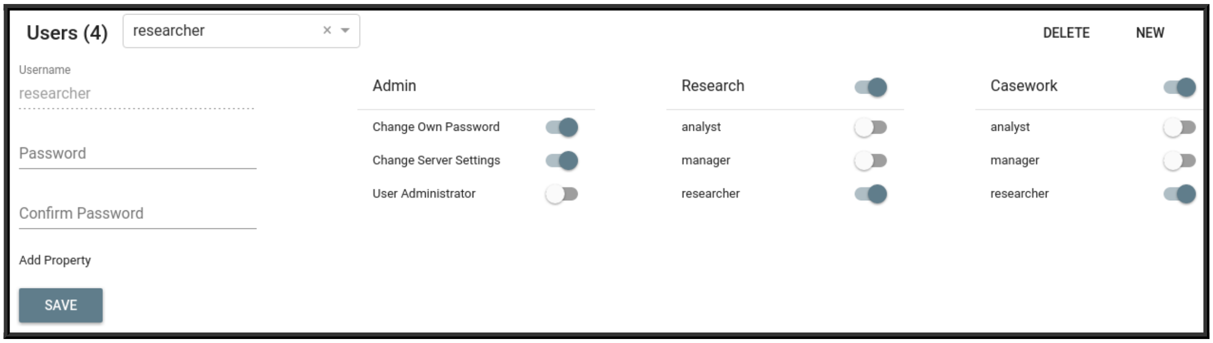
USERS Tab in Admin mode

This interface allows users to be added, deleted, or modified including changing the user password.

User management in MaSTR is based on the concept of users being assigned to 'groups'. Each group consists of several permissions, and a user's potential access in a given database is limited to the combined access of all the groups they belong to.

Each user can be assigned one or more groups for both the 'work' database and the 'research' database. Users may also be activated/deactivated in those databases using the toggle next to the database name, which acts as enabling/disabling their membership in a built-in `User` group. This group allows login and viewing of any information, so deactivated users are unable to login to that database.

There are 3 possible groups which can be assigned in the admin database, controlling user management and server settings. All permissions are granted in the 'demo' database, but the demo database is cleared on logout.



Users (4) researcher

Username: researcher

Password

Confirm Password

Add Property

Admin		Research	●	Casework	●
Change Own Password	●	analyst	○	analyst	○
Change Server Settings	●	manager	○	manager	○
User Administrator	○	researcher	●	researcher	●

Default group assignments for the 'researcher' user account

Admin Groups

The three groups in the admin database function like specific permissions.

1. **Change Own Password** - The user has the ability to change their own password
2. **Change Server Settings** - The user has the ability to change settings on the server. Most of these are related to report formatting.
3. **User Administrator** - The user has the ability to access the admin interface, including modifying the permissions and passwords of their

own account and any other users (aside from *root*).

Research and Casework Groups

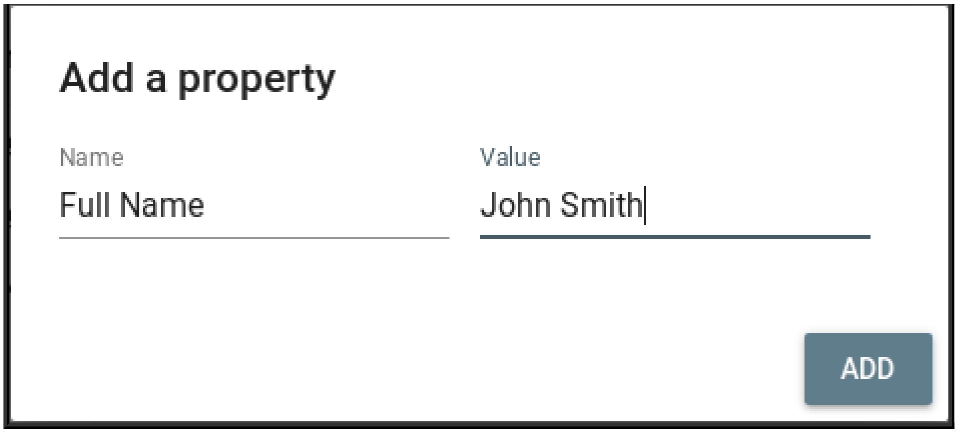
There are several built-in groups controlling permissions in the `research` and `casework` databases. Groups in these databases can be deleted or modified (given different permissions) and new groups can be created by the root user.

1. **Manager** - grants access to all permissions in the database that it is granted for.
2. **Researcher** - grants access to all permissions except deleting jobs
3. **Analyst** - grants access to all permissions needed to run and view jobs, but doesn't allow upload, modification, or creation of associated files (panels, frequencies, models, and SOPs).

> **Note**
>
> The built-in groups are included as basic defaults. They can be modified or replaced entirely. Any user without an assigned group will have basic login and viewing permissions as part of the built-in `User` group unless their access to the database is disabled.

Properties

Miscellaneous information can be added to each user with the "Add property" button. This information is for reference only, and is not used directly by the software.



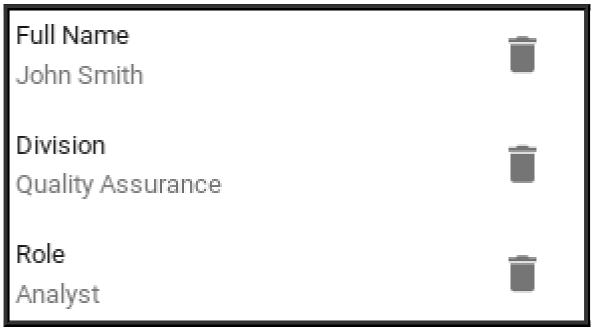
Add a property

Name: Full Name

Value: John Smith

Add property dialog

The information is listed when viewing that user.



Full Name
John Smith

Division
Quality Assurance

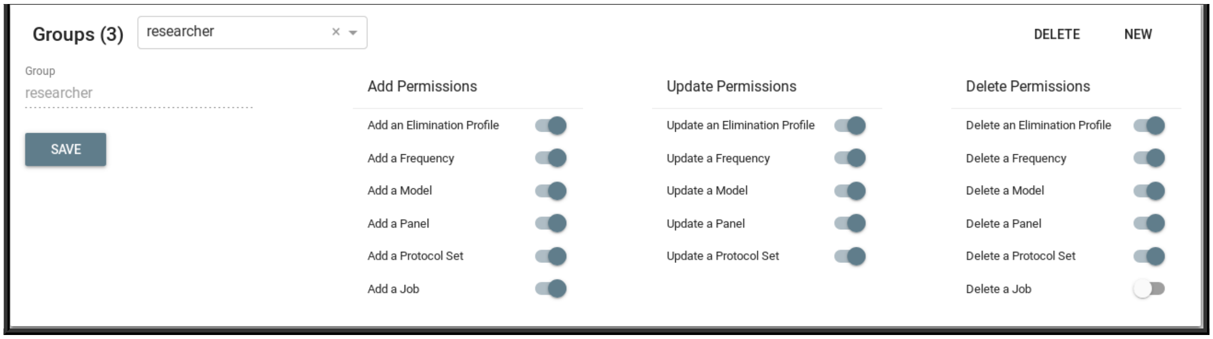
Role
Analyst

Example user information

GROUPS Tab

Each group can have individual roles toggled on or off. These roles are grouped into:

- *Add Permissions* - upload and create new files/jobs
- *Update Permissions* - modify existing files
- *Delete Permissions* - remove existing files/jobs



Groups (3) researcher × ▾ DELETE NEW

Group
researcher

SAVE

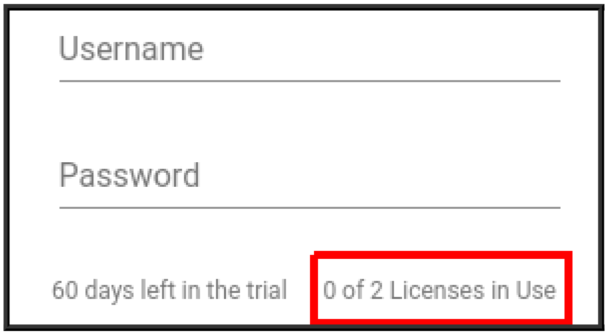
Add Permissions		Update Permissions		Delete Permissions	
Add an Elimination Profile	☑	Update an Elimination Profile	☑	Delete an Elimination Profile	☑
Add a Frequency	☑	Update a Frequency	☑	Delete a Frequency	☑
Add a Model	☑	Update a Model	☑	Delete a Model	☑
Add a Panel	☑	Update a Panel	☑	Delete a Panel	☑
Add a Protocol Set	☑	Update a Protocol Set	☑	Delete a Protocol Set	☑
Add a Job	☑			Delete a Job	□

Default group roles for the 'researcher' group. All permissions are enabled except for deleting jobs

Sessions

Login

The login screen will display the number of available licenses. A license is in-use when a user is logged in to any mode other than `admin`. If all licenses are in-use, it is not possible to login until a user logs out, making a license available.



Username

Password

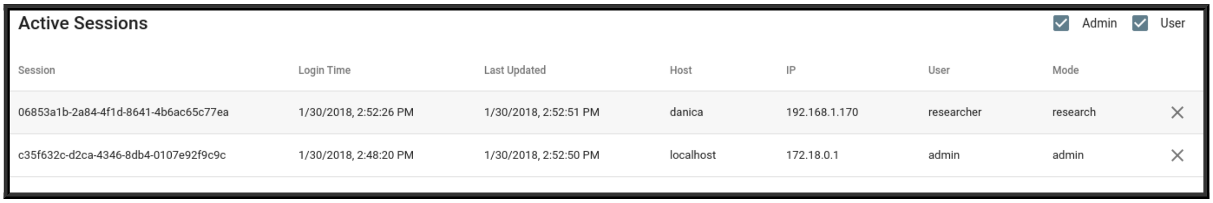
60 days left in the trial

0 of 2 Licenses in Use

Login screen showing 0 of 2 licenses in use

Managing Sessions

The admin interface includes a `SESSIONS` tab that shows the current active sessions. Sessions may be killed using the 'X' next to the session information. When this happens the user of that session is signed out. The display of `Admin` (doesn't use a license) or `User` sessions may be toggled.



Active Sessions

☑ Admin ☑ User

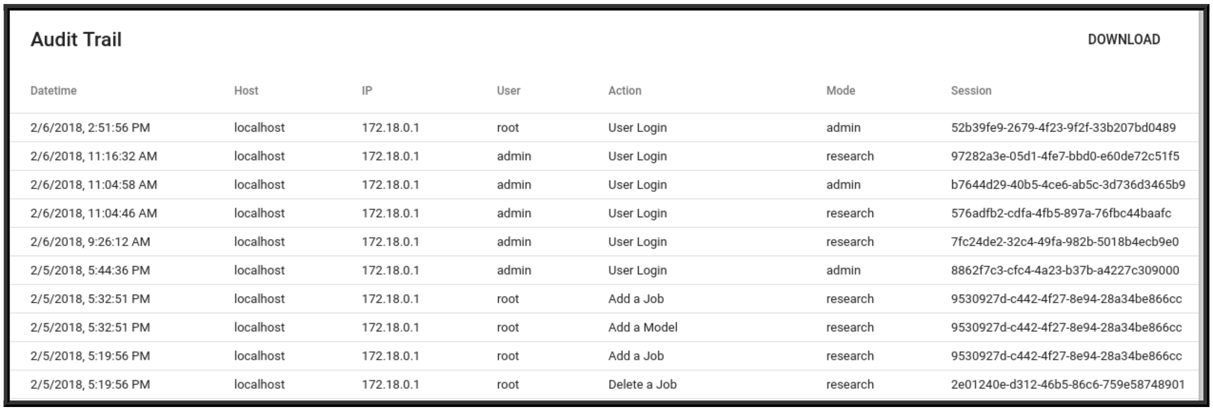
Session	Login Time	Last Updated	Host	IP	User	Mode	
06853a1b-2a84-4f1d-8641-4b6ac65c77ea	1/30/2018, 2:52:26 PM	1/30/2018, 2:52:51 PM	danica	192.168.1.170	researcher	research	✕
c35f632c-d2ca-4346-8db4-0107e92f9c9c	1/30/2018, 2:48:20 PM	1/30/2018, 2:52:50 PM	localhost	172.18.0.1	admin	admin	✕

The `SESSIONS` tab in admin mode showing two active sessions- an admin session from the current computer and the 'researcher' user logged in to the research db from another computer

Audit Trail

MaSTR records the username and time of every API request that may result in a change to information in the database.

These records can be viewed and downloaded on the *LOG* tab of the admin interface.



Audit Trail

Datetime	Host	IP	User	Action	Mode	Session
2/6/2018, 2:51:56 PM	localhost	172.18.0.1	root	User Login	admin	52b39fe9-2679-4f23-9f2f-33b207bd0489
2/6/2018, 11:16:32 AM	localhost	172.18.0.1	admin	User Login	research	97282a3e-05d1-4fe7-bbd0-e60de72c51f5
2/6/2018, 11:04:58 AM	localhost	172.18.0.1	admin	User Login	admin	b7644d29-40b5-4ce6-ab5c-3d736d3465b9
2/6/2018, 11:04:46 AM	localhost	172.18.0.1	admin	User Login	research	576adfb2-cdfa-4fb5-897a-76fbc44baafc
2/6/2018, 9:26:12 AM	localhost	172.18.0.1	admin	User Login	research	7fc24de2-32c4-49fa-982b-5018b4ecb9e0
2/5/2018, 5:44:36 PM	localhost	172.18.0.1	admin	User Login	admin	8862f7c3-cfc4-4a23-b37b-a4227c309000
2/5/2018, 5:32:51 PM	localhost	172.18.0.1	root	Add a Job	research	9530927d-c442-4f27-8e94-28a34be866cc
2/5/2018, 5:32:51 PM	localhost	172.18.0.1	root	Add a Model	research	9530927d-c442-4f27-8e94-28a34be866cc
2/5/2018, 5:19:56 PM	localhost	172.18.0.1	root	Add a Job	research	9530927d-c442-4f27-8e94-28a34be866cc
2/5/2018, 5:19:56 PM	localhost	172.18.0.1	root	Delete a Job	research	2e01240e-d312-46b5-86c6-759e58748901

Example audit trail logs

Probabilistic Mixture Analysis

What is Probablistic Mixture Analysis?

Probablistic Mixture Analysis is a fully-continuous bayesian approach to resolving complex mixture data.

What is being calculated?

Likelihood Ratio

$$\frac{P_{Prosecution}}{P_{Defense}} = \frac{P_{Suspect\ is\ a\ Contributor}}{P_{Contributor\ is\ a\ random\ person}}$$

where:

$$\frac{P_{Prosecution}}{P_{Defense}} = \frac{\sum_{Genotype\ Sets\ Including\ Suspect} P(Genotype\ Set \mid Data)}{\sum_{Any\ Possible\ Genotype\ Set} P(Genotype\ Set \mid Data)}$$

What is P(Genotype Set | Data)?

Bayes Theorem states:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

This is used to determine the probability of a given set of contributors given the mixture data:

$$P(Genotypes \mid Data) = \frac{P(Data \mid Genotypes)\, P(Genotypes)}{P(Data)}$$

This can be further broken down. If we can describe a model that specifies expected peak heights based on some input set of genotypes (such as 14, 14 and 13, 14) and several parameter values (contributor ratio, template DNA amount, etc), we can compare the expected peak height to the observed data in order to assign a weight to those genotypes and parameter values. This weight is based on the amount of variance we expect to see from run to run- how similar the peak heights would be if we could use the same sample in the machine multiple times.

$$P(Data \mid Genotypes) = P(Data \mid Expected\ Heights, Variance)$$

$$P(Genotype\ Set) = Population\ Frequency\ Probability$$

$$P(Data) = \int_{Model\ Parameters} P(Data|Model\ Parameters)$$

Note

Calculating P(Data) - the integral over all possible model parameters - is extremely difficult. Luckily it can be approximated with high accuracy using **Markov Chain Monte Carlo (MCMC)**, which will be explained later on.

MCMC

What is *MCMC*?

MCMC Stands for **Markov Chain Monte Carlo**. Definitions for a few related concepts may be useful:

- Markov Property - A stochastic process where the conditional probability of any future states depends only on the present state and not on past states
- Markov Process - A stochastic process that satisfies the Markov Property
- Markov Chain - A process that occurs in a series of time-steps (iterations) where a random choice is made in each step

Markov Chain Monte Carlo is an algorithm used to optimize hidden parameter values in a complex model. This is done in a few steps:

1. Pick some random values based on the input distributions
    - Ideally these input distributions are as unbiased as possible.
    - As an example, the software randomly picks a ratio value for each contributor in the mixture
2. Calculate a probability
    - The Model is used to generate hypothetical peak heights based on the selected alleles, ratio values, and other sampled variables.
    - The hypothetical peaks are compared to the actual signal data. A probability is calculated based on the expected variance from run to run (as calculated in the `Protocol Set`).

3. Accept or reject the new values.
    - If the new values are more likely than the previous values in the chain, acceptance is more likely.
4. Repeat for some number of iterations

This process results in a chain (or "trace") of values. It can be likened to a game of "Hot or Cold" where a "searcher" is trying to find a hidden object.

1. The searcher takes a random step.
2. The observer tells them if they are "hotter" (closer to the object) or "colder" (farther away).
3. The searcher makes a decision.
    - If they are colder, it makes sense to undo the previous step and try a new random direction.
    - If they are warmer, it makes sense to keep that step and make another (hopefully also warm) step.
4. This process repeats until the searcher finds the object.

The distinction in this case is that there is no final known location. The searcher will tend to spend more time close to the correct position, and therefore the position of the searcher in the room over time gives a probability distribution of the location of the object.

Terms

Iterations - The number of attempted steps. Using more iterations will make the final estimate more consistent and less prone to random variations in the likelihood ratios. If the number of iterations is low, the process may get "lucky" and find a good "spot" or may be unlucky and never find a good one. Using a larger number of iterations can make the search process give more consistent results.

Burn-In - Early in the process the probability is low and the variables aren't likely to be accurate. The final distribution of variables would get more accurate over a large number of iterations as the more recent iterations outnumber the early ones. This can be improved further by ignoring the results of the initial iterations/steps.

Thinning - The 'mixing' of the chain (how much movement occurs) can be improved by ignoring every other (or every 3rd, 4th, etc) step. Doing so requries that the total number of iterations is increased, which increases runtime.

Chains

Sampling Methods

There are a few different methods for how the software chooses which alleles or variable values to test.

Allele Step - Sampling alleles in the MCMC process

Subset Merge - An MCMC method that tries to move to a new genotype list each step by modifying the current list slightly each iteration. Almost as fast as uniform, and thus recommended for most use cases.

Uniform - Generates a new genotype list each iteration by assigning each allele in every contributor and tries to move there. Fast, but will potentially move very slowly if there are many genotype lists, as most will have a very low probability. Probably should only be used for a 2 or 3 contributors.

Continuous Step

Metropolis - A simple, fast MCMC method for the continuous variables. Recommended for most uses.

Slice - A more advanced method that is able to move around more in each iteration at the expense of more computation, which makes it slower than Metropolis.

Probability Calculation from MCMC Results

After calculating expected peak heights, they are compared to observed peak heights in order to generate a probability. The overall probability is calculated as the product of the probability of each peak height comparison.

The goal is to get:

$$P(Data \mid Genotypes) = P(Data \mid Expected\ Heights, Variance)$$

To fit into the original formulation:

$$P(Genotypes \mid Data) = \frac{P(Data \mid Genotypes)\, P(Genotypes)}{P(Data)}$$

There are 3 possible cases:

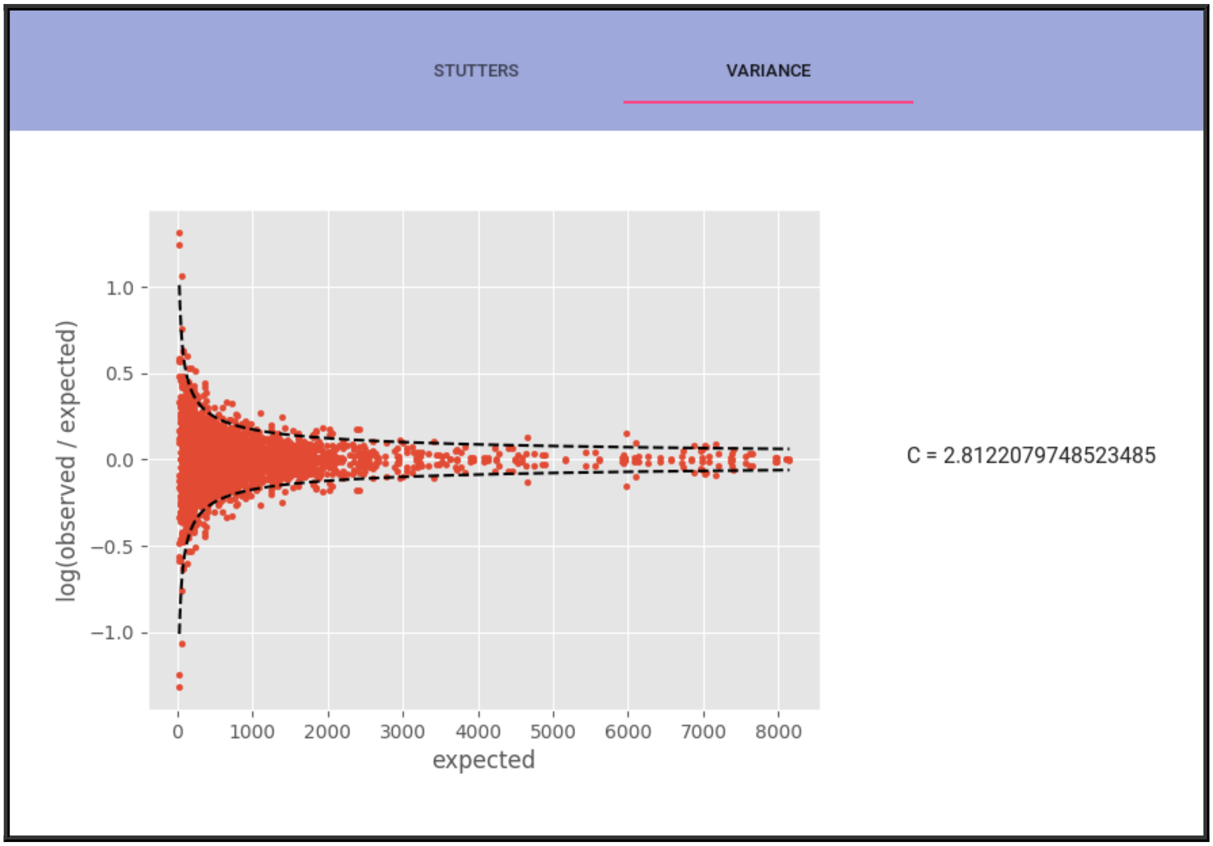
Expected and Observed

When a peak is both expected and observed, the heights of each are compared and a probability is generated. The variance (calculated in the protocol set) is an important part of this calculation.

$$P(Data \mid Expected, OtherVariables) = LogNormal_{(}\mu = Expected,$$

Variance

Even in identical conditions, the peak heights of of a sample won't be identical across multiple runs. The variance is a measurement of how much noise is expected from run to run. After expected peak heights are calculated this information is used to judge the relative weight of that prediction.



C = 2.8122079748523485

Variance plot showing the fitted variance measurement

As the peak height increases, a lower log-fold change is expected between the expected height and the actual height.

Drop-out

Drop-out is when a peak is expected, but not observed. When generating genotypes it is possible to include an expected *Q* allele which represents any allele that was not called in the signal data. Since there is no single observed height, the height is integrated over all values from 0 to the analytic threshold.

$$P(Data \mid Expected, OtherVariables) = \int_0^{AT} (P_{Observed} | P_{Expected,Othe}$$

Warning

Drop-out of two alleles in the same genotype is not considered as a possibility when generating potential sets of genotypes.

Drop-in

Drop-in occurs when a peak is observed, but not expected. Models include a "Drop-in Coefficient" λ which determines the likelihood of a drop-in peak based on the peak height.

$$P(Data \mid Expected, OtherVariables) = \lambda \exp(-\lambda Observed)$$

Lowering the drop-in coefficient will make genotype sets with drop-in alleles more likely.

Coancestry Adjustment

There are 3 possible methods for adjustments to the population frequency values based on coancestry assumptions:

None (HWE)

$$\text{Homozygous} = p^2$$

$$\text{Heterozygous} = 2pq$$

NRCII Recommendation 4.1

$$\text{Homozygous} = p^2 + p(1-p)\theta$$

$$\text{Heterozygous} = 2pq$$

NRCII Recommendation 4.2

$$\text{Homozygous} = \frac{(2\theta + (1-\theta)p)(3\theta + (1-\theta)p)}{(1+\theta)(1+2\theta)}$$

$$\text{Heterozygous} = \frac{(\theta + (1-\theta)p)(\theta + (1-\theta)q)}{(1+\theta)(1+2\theta)}$$

Elimination Database

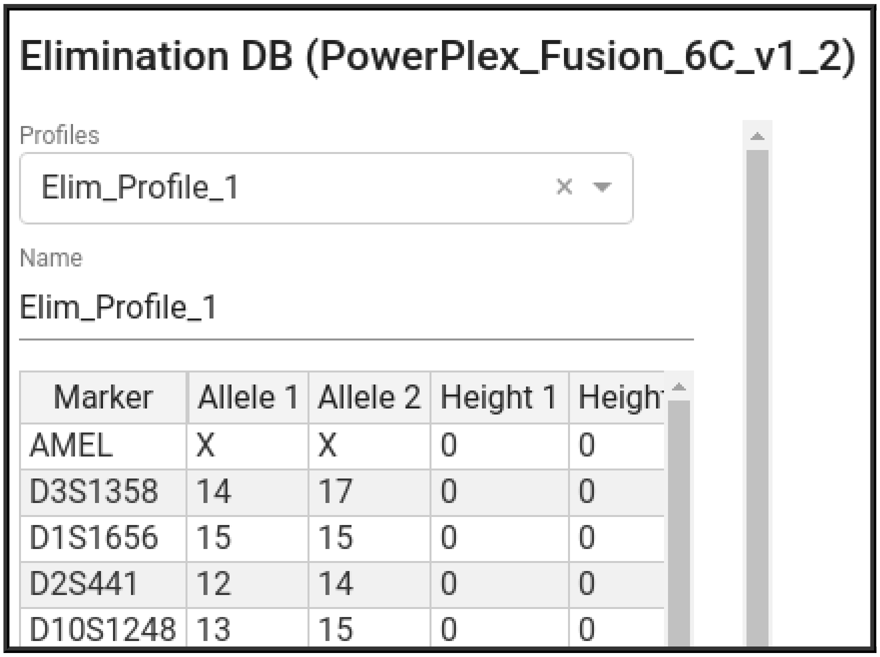
The elimination database may be used to automatically flag any potential contamination of samples.

Profiles

Profiles are associated with a panel and may be viewed/modified using the *ELIMINATION DB* button on the panel view. Profiles may be imported from files:

```
Sample Name Marker  Allele#1    Allele#2
PersonA D8S1179 13  12
PersonA D21S11  28  31
PersonA D7S820  11  11
PersonA CSF1PO  10  10
PersonA D3S1358 15  18
PersonA TH01    9   11
PersonB D8S1179 12  12
PersonB D21S11  30  31
PersonB D7S820  9   11
PersonB CSF1PO  10  12
PersonB D3S1358 15  18
PersonB TH01    7   8
PersonC D8S1179 13  16
PersonC D21S11  28  32.2
PersonC D7S820  7   11
PersonC CSF1PO  7   11
PersonC D3S1358 16  16
PersonC TH01    7   7
```

If a name isn't included, the filename is used. If a name is included, multiple profiles may be included in a single file. Profile names must be unique in the elimination db for the associated panel.



Elimination DB (PowerPlex_Fusion_6C_v1_2)

Profiles

Elim_Profile_1

Name

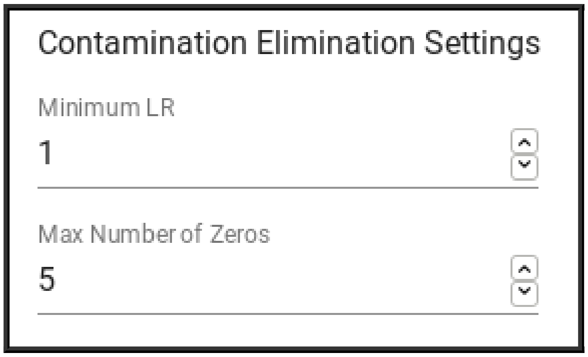
Elim_Profile_1

Marker	Allele 1	Allele 2	Height 1	Heigh
AMEL	X	X	0	0
D3S1358	14	17	0	0
D1S1656	15	15	0	0
D2S441	12	14	0	0
D10S1248	13	15	0	0

An elimination profile with an automatically generated name

Filters

All elimination profiles associated with a panel are tested when a job is run. Any profiles passing filter requirements (maximum number of 0-lr markers and minimum overall LR) are flagged in the job results. The filters are set in the protocol set view.



Contamination Elimination Settings

Minimum LR

1

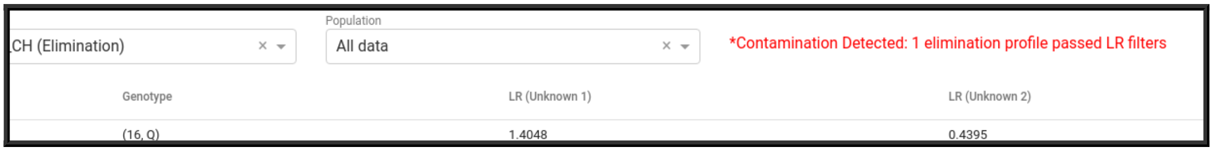
Max Number of Zeros

5

Elimination filters set in the protocol set

Reporting Results

Red text in the job results indicates that elimination profiles were passed the filters, and the elimination profile results can be viewed by selecting the profile in the drop-down.



CH (Elimination)

Population: All data

*Contamination Detected: 1 elimination profile passed LR filters

Genotype	LR (Unknown 1)	LR (Unknown 2)
(16, Q)	1.4048	0.4395

Flagged elimination profile

Calculating Peak Heights

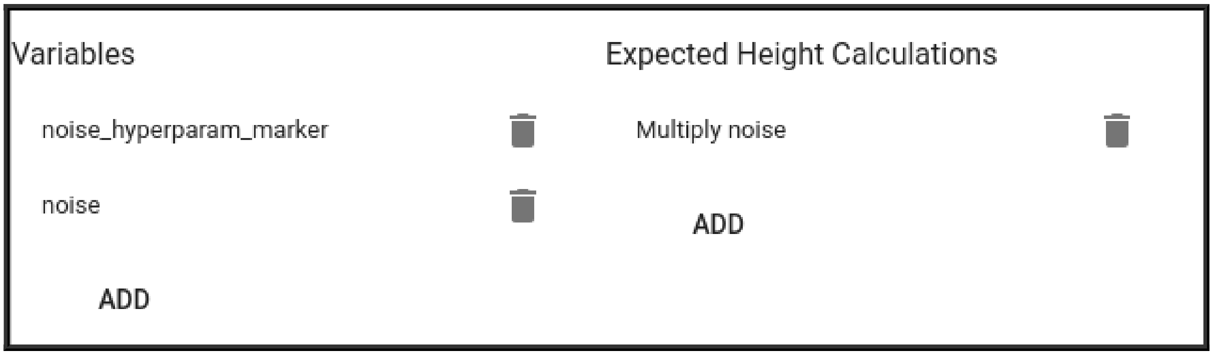
Model Structure

MaSTR's calculation of expected peak heights has several steps:

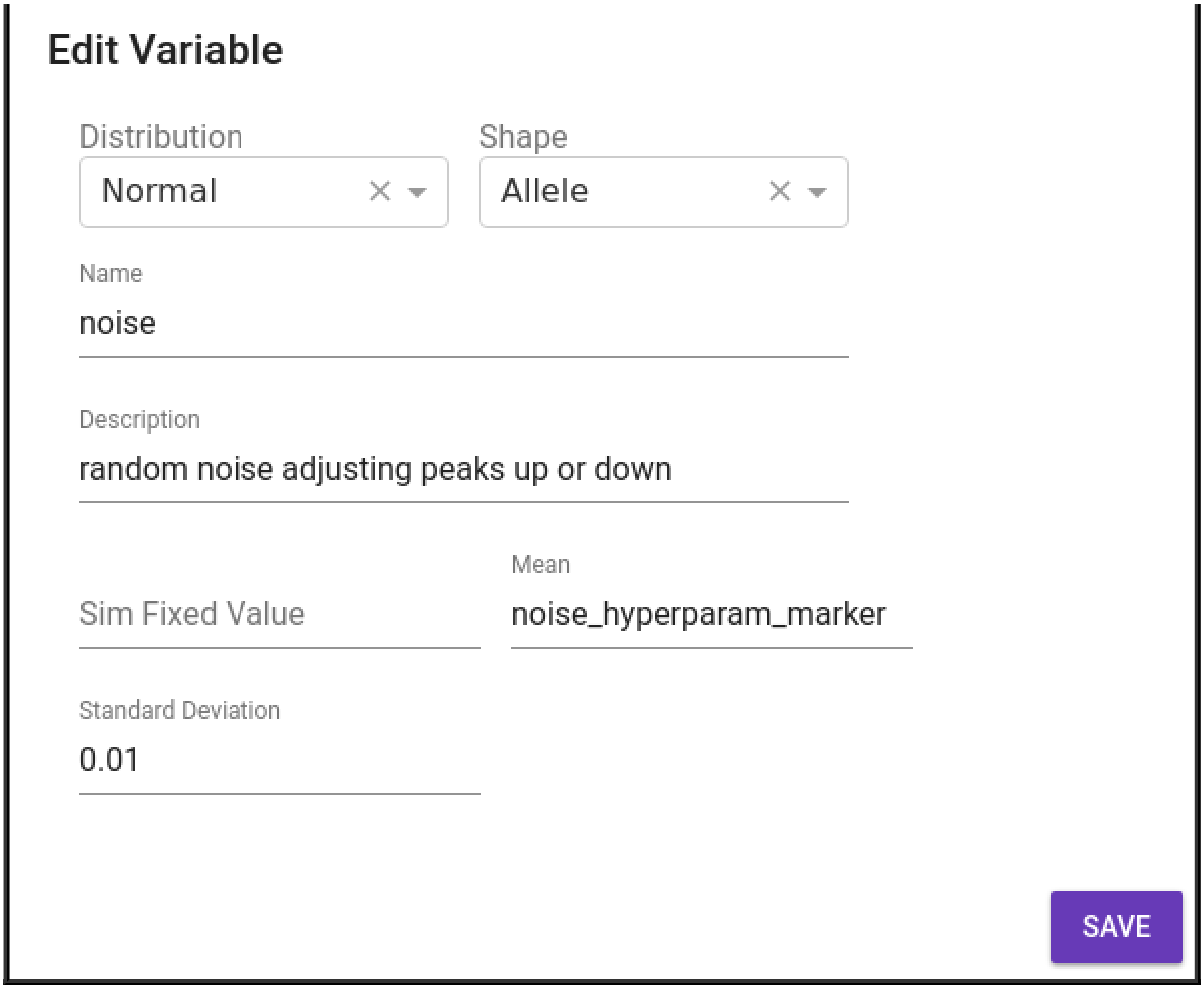
- Start with some suitable mixture of genotypes (a value of 1 per allele)
    - Assume contributors are 14,13 and 14,10
    - 14 = 2, 13 = 1, 10 = 1, other alleles in the panel = 0
- Apply a contributor ratio
    - Assume 75% : 25%
    - 14 = 1, 13 = 0.75, 10 = 0.25
- Apply the 'amount' values
    - One overall *amount* multiplied by *A* for each marker (see degradation)
    - Calculate 4000 total peak height (2000 per allele) with an *A* value of 0.9 for this marker = 1800 per allele
    - 14 = 1800, 13 = 1350, 10 = 450
- Apply Degradation
    - A *D* value for each contributor lowers the expected height of larger alleles for that contributor (see degradation)
- Apply any custom addition, subtraction, multiplication, or division of custom model variables as defined in the **Model**
- Apply any stutter specified in the protocol set
    - 14 > 13 = 0.1, 13 > 12 = 0.08

- 14 = 1620, 13 = 1422, 12 = 108, 10 = 450

Custom **Model Variables** may work in a hierarchical way. For instance, a per-marker "degradation" variable may be used as a prior in specifying a per-contributor variable that modifies the expected heights of each contributor. These additional variables may be used to add, subtract, multiply, or divide the peak heights before applying the stutter calculation.



Model Variables



Edit Variable

Distribution: Normal

Shape: Allele

Name: noise

Description: random noise adjusting peaks up or down

Sim Fixed Value

Mean: noise_hyperparam_marker

Standard Deviation: 0.01

A per-allele 'noise' variable defined using a normal distribution. The mean value of the distribution is another variable called 'noise_hyperparam_marker', which represents one random value per marker sampled from a uniform distribution between 0.95 and 1.05. The standard deviation is 0.01.

Sampler Warmup / Preprocessing

Sampler warmup may be enabled in the model settings. This runs a simplified version of the mcmc process in order to select relatively likely genotypes for unknown contributors which are then used as the starting point of the main MCMC process.

The only unknown variable in this simplified process is the contributor allele selection- the ratio is divided evenly amongst contributors, and other variables are held at default starting values.

Without the warmup process it is possible that the main MCMC process may become stuck with some poor values for some parameters.

Note

When running multiple chains in the main MCMC process, each chain will use the same starting values for allele selection when preprocessing is enabled.

Degradation

What is Degradation?

When a sample becomes degraded, the template DNA becomes fragmented. This has a deleterious effect on the amplification of larger fragments. This effect should be accounted for when generating expected peak heights.

How does MaSTR account for it?

The model settings include a degradation drop-down with three possible settings:

- *No Adjustment* - The potential effects of degradation are ignored.
- *Linear* - Degradation is assumed to occur with a linear relationship to allele size
- *Exponential* - Degradation is assumed to occur with an exponential relationship to allele size.

Degradation Plots

Several variables are accessible on the **DEGRADATION PLOTS** tab of the job results view.

A

- One value for each marker. One marker is set to 1.0 and all others are realtive to that marker.
- Some markers may amplify more or less efficiently than others. This value allows for each marker to have a different peak height given some number of copies of an allele, all other things being equal.

mu_A

- The estimated average value of *A*

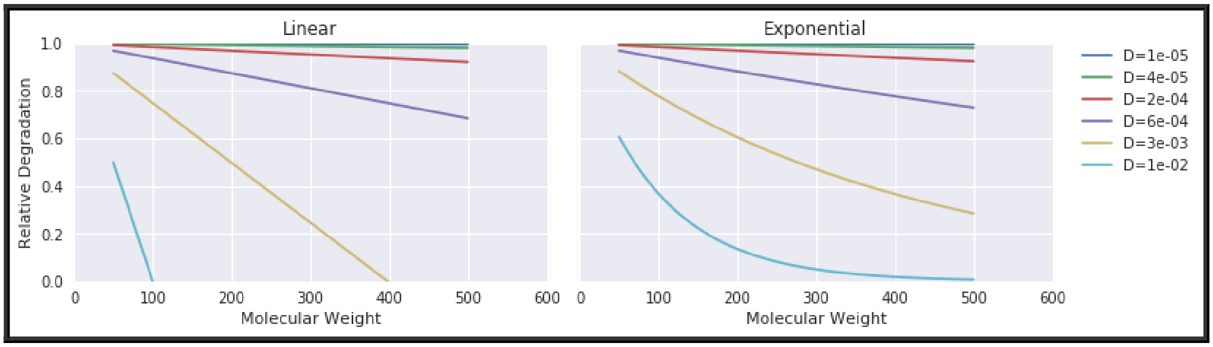
amount

- The absolute peak height value that scales the peaks heights. This causes the expected peak heights to have real values (645, 992, 87, etc) instead of relative values (0.5375, 0.8267, 0.0725).

D

- One value per contributor.
- Degradation parameter used to calculate degradation for a given molecular weight (either linear or exponential)

The following plot shows degradation values across molecular weights for example values of *D*.



Protocol Set

Several values related to degradation are part of the protocol set.

- Linear Variance
- Exponential Variance
- Linear Sigma
- Exponential Sigma

> ⚠ **Warning**
>
> Currently these values are hard-coded to appropriate, widely-applicable values based on real data. In the future they will be calculated as part of the protocol set process.

Sigma values are a measure of the variance in the *A* values. The *A* scaling factor should vary between markers, but this sigma value limits how much it should vary.

When degradation is used, the Linear Variance or Exponential Variance is used as *C* in the probability calculation.

Stutter

Calculation

Stutter ratios are calculated as part of the Protocol Set using single-source data.

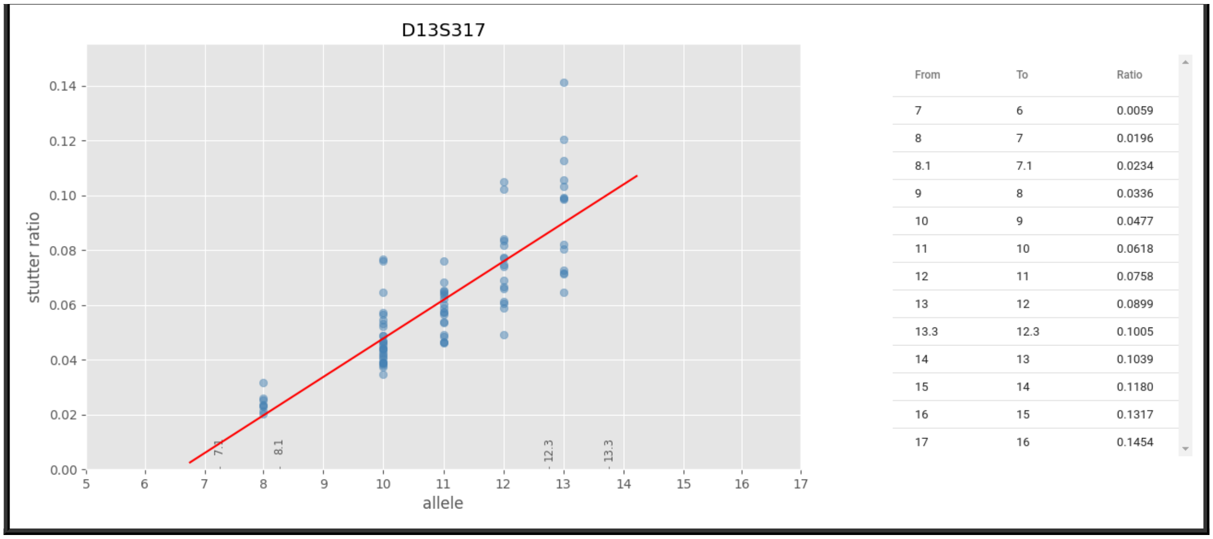
1. Collect stutter ratio values

For each sample and each marker in the panel, the stutter ratios are measured by dividing the height of the stutter peak at the minus-1 repeat position by the height of the peak from the contributor's allele.

> **Note**
>
> If a contributor's genotype is (`A-1`, `A`) at a marker, their samples will not be used for calculating stutters at this marker, because the `A-1` peak is partly due to stutter and partly due to the genotype

2. Apply a linear regression

For each marker a linear regression is applied to this stutter data, to estimate the stutter ratio as a function of allele length. This will allow the estimation of the stutter ratio for any allele.



From	To	Ratio
7	6	0.0059
8	7	0.0196
8.1	7.1	0.0234
9	8	0.0336
10	9	0.0477
11	10	0.0618
12	11	0.0758
13	12	0.0899
13.3	12.3	0.1005
14	13	0.1039
15	14	0.1180
16	15	0.1317
17	16	0.1454

Linear-fit of stutter for D13S317. Blue dots show the raw values and the red line shows the fitted equation

Application of Stutter

When enabled, stutter ratios for each allele are used to move some of the expected peak height from each allele to it's stutter position.

Additional methods

Other stutter models are in-development, such as an LUS (longest uninterrupted stretch) model and stutter at positions other than "-1".

Model Variables

Essential Concept

Model Variables may be defined as part of a *Model* in an attempt to better represent the underlying processes which generate allele peak height measurements from DNA. Values are randomly sampled from specified distributions and optimal values for each variable are found using the MCMC process. There is a trade-off between model complexity (more model variables result in slower runtimes and higher memory usage) and model accuracy (adding some model variables may improve the model's accuracy).

Defining Model Variables

Each model variable definition includes:

- *Distribution* - There are several choices of distribution built into the software and listed below.
- *Shape* - The number of values that are sampled from the distribution.
  - Marker - one per marker
  - Model - one value for the entire model
  - Allele - one per allele
  - Contributor - one per contributor
- *Name* - A unique name for the model variable
- *Description* - optional description of what the variable represents

The sampled values may then be added to the *Expected Height Calculation* via multiplication, division, addition, or subtraction of the values from the calculated pre-stutter signal.

Hierarchical Models

Model Variables may be used as input parameters in subsequent model variables, creating hierarchical models. Shapes are automatically broadcast as needed. For example, an `allele` variable with a `marker` prior variable will generate a value for each allele using the prior value for the corresponding marker.

Distributions

Fixed

A fixed distribution uses a single value instead of sampling from a distribution. This is useful for creating hierarchical variables with different shapes. These model variables are not shown in the results since their values do not change throughout the trace.

Normal

Uniform

Bounded Normal

Symmetric Dirichilet

Beta

Gamma

Bernoulli

Lognormal

Exponential

Open Source Libraries

The software utilizes many open source software packages and libraries:

Server

- Python
- MongoDB
- Crossbar
- Python Libraries
    - MKDocs
    - Numpy
    - Scipy
    - Matplotlib
    - Pandas
    - PyJWT
    - PyMongo
    - PyMC3
    - Theano
    - Hug
    - Marshmallow
    - Click
    - Autobahn

Client

- NodeJS
- React
- Material-UI
- D3
- Plotly.js
- Autobahn

The source code for MaSTR is available to customers on request.