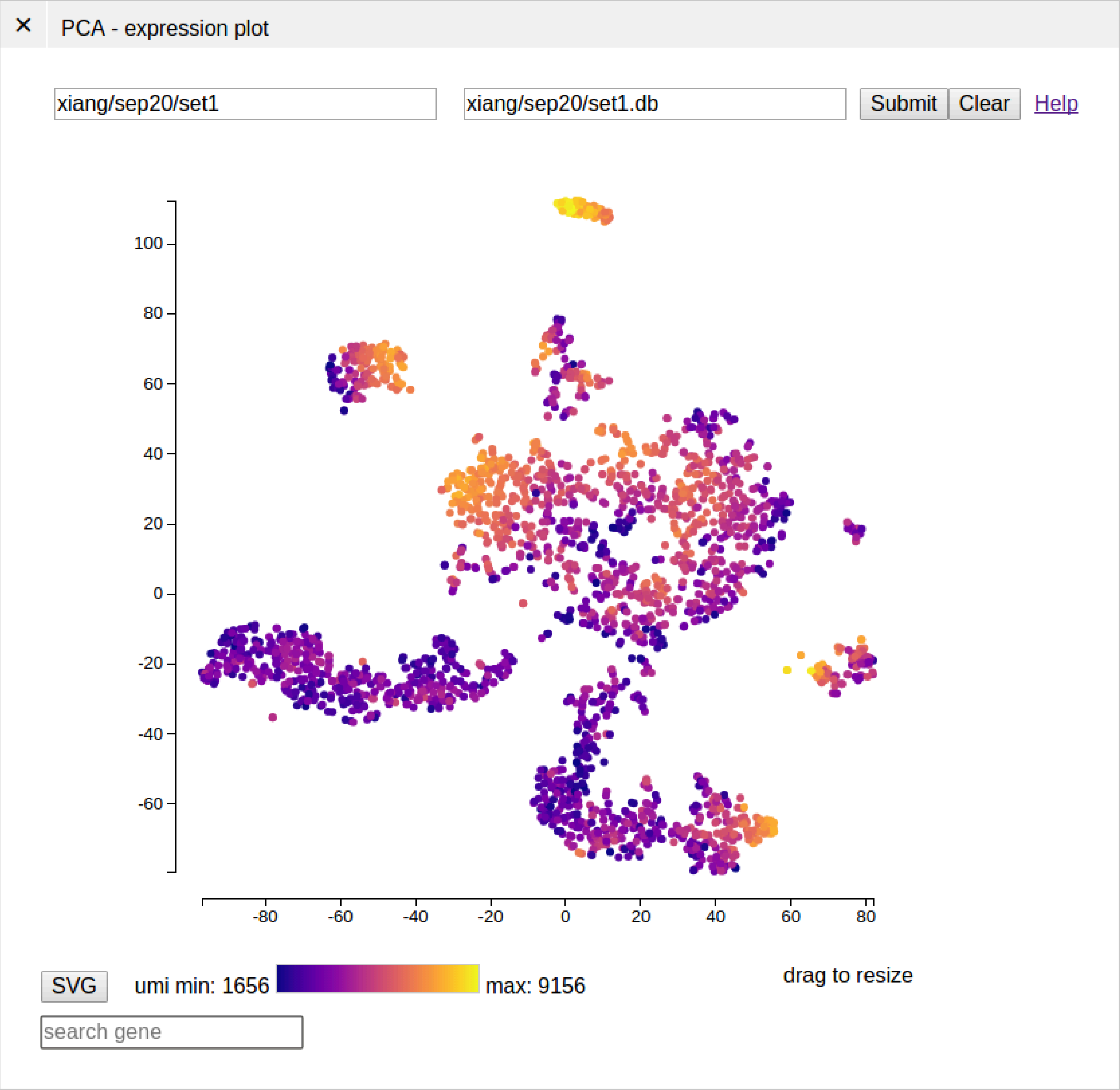
PCA - expression plot



xiang/sep20/set1

xiang/sep20/set1.db

This function requires two type of data for input:

1. PCA/t-SNE vector of many samples, used for drawing samples a 2-D scatterplot.
    a. Data in the form of a text file
2. Gene expression levels (or other numerical values) of the same set of samples from the PCA plot.
    a. Data in the form of a SQLite database

Both the vector file and the db file should be hosted on the ProteinPaint server, under the <TP> directory. To use this function, provide the relative path of the vector file and the db file. When the scatterplot is rendered, enter gene name to retrieve gene expression value for all samples, and overlay it with the scatterplot.

File format:

Vector file has 3 columns:

1. Sample name
2. X-axis coordinate
3. Y-axis coordinate

Database file has 3 columns:

1. Gene name, or any string as the key for associating numerical values
    a. **Must be lower-case!**
2. Sample name, must exactly match with that in the vector file
3. Expression value for this gene

To prepare the SQLite database:

Get sqlite3 precompiled binary from https://www.sqlite.org/download.html

Do following:

```
$ cat > db.sql
drop table if exists data;
CREATE TABLE data (
gene varchar(255) not null,
sample varchar(255) not null,
value float not null
);

.mode tabs
.import /path/to/expressiondata.txt data
CREATE INDEX data_gene on data (gene);

$ sqlite [filename].db < db.sql
```

This will generate the db file called “filename.db”.

The table name as well as schema are currently hardcoded.

Create the index after loading data to speed up.