# Mash comparison of mock data

The mock community data is used as a baseline data for testing the analysis pipeline and evaluating the cut-off value for the maximum mash distance for distinguishing species.

```
In[ ]:= Clear["Global`*"];
        SetDirectory[NotebookDirectory[]];
```

```
In[ ]:= distanceCsv = Import["data/mock-mash/dist.csv.gz"];
        distanceMat = distanceCsv[[2 ;;]];
        cellNames = ToString /@ distanceCsv[[1]];
        species = ToString[#[[1]] → #[[2]] & /@ Import["data/mock-mash/species.tsv"]];
```

The four species: BS (*Bacillus subtilis*), EC (*Escherichia coli*), KP (*Klebsiella pneumoniae*), SA (*Staphylococcus aureus*)

```
In[ ]:= bsCells = Select[species, #[[2]] == "BS" &][[All, 1]];
        ecCells = Select[species, #[[2]] == "EC" &][[All, 1]];
        kpCells = Select[species, #[[2]] == "KP" &][[All, 1]];
        saCells = Select[species, #[[2]] == "SA" &][[All, 1]];
```

The matrix indices of the species.

```
In[ ]:= bsIndices = Flatten[Position[cellNames, #] & /@ bsCells];
        ecIndices = Flatten[Position[cellNames, #] & /@ ecCells];
        kpIndices = Flatten[Position[cellNames, #] & /@ kpCells];
        saIndices = Flatten[Position[cellNames, #] & /@ saCells];
```

```
In[ ]:= FlattenWithoutDiagonal[submat_List] := Module[{n},
          n = Length[submat];
          Flatten[Table[Delete[submat[[i]], i], {i, n}]]
        ];
```

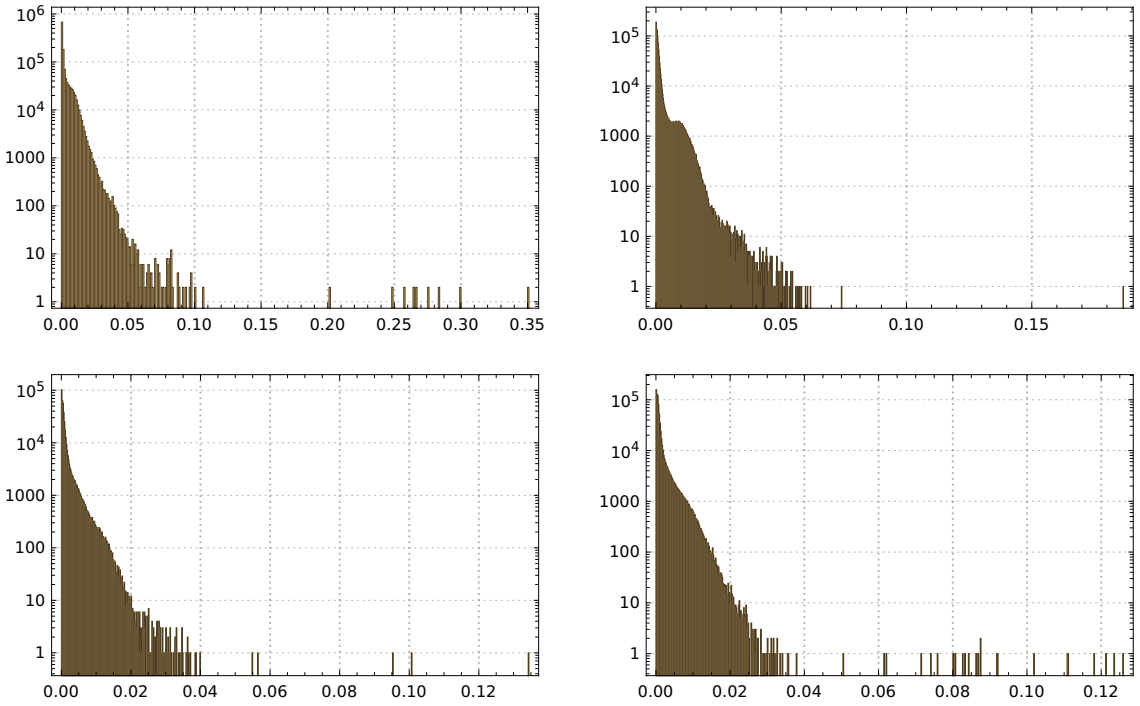
The distances on the diagonal (comparing to self) should be crossed out.

```
In[ ]:= bsSimilarities = FlattenWithoutDiagonal[distanceMat[[bsIndices, bsIndices]]];
        ecSimilarities = FlattenWithoutDiagonal[distanceMat[[ecIndices, ecIndices]]];
        kpSimilarities = FlattenWithoutDiagonal[distanceMat[[kpIndices, kpIndices]]];
        saSimilarities = FlattenWithoutDiagonal[distanceMat[[saIndices, saIndices]]];
```

## BS vs BS, EC, KP, SA

In[ ]:=
```
GraphicsGrid[ArrayReshape[
  Histogram[#, ScalingFunctions → "Log", PlotTheme → "Detailed", PlotRange → Full] & /@ {
    bsSimilarities,
    Flatten[distanceMat[[bsIndices, ecIndices]]],
    Flatten[distanceMat[[bsIndices, kpIndices]]],
    Flatten[distanceMat[[bsIndices, saIndices]]]
   }, {2, 2}]]
```

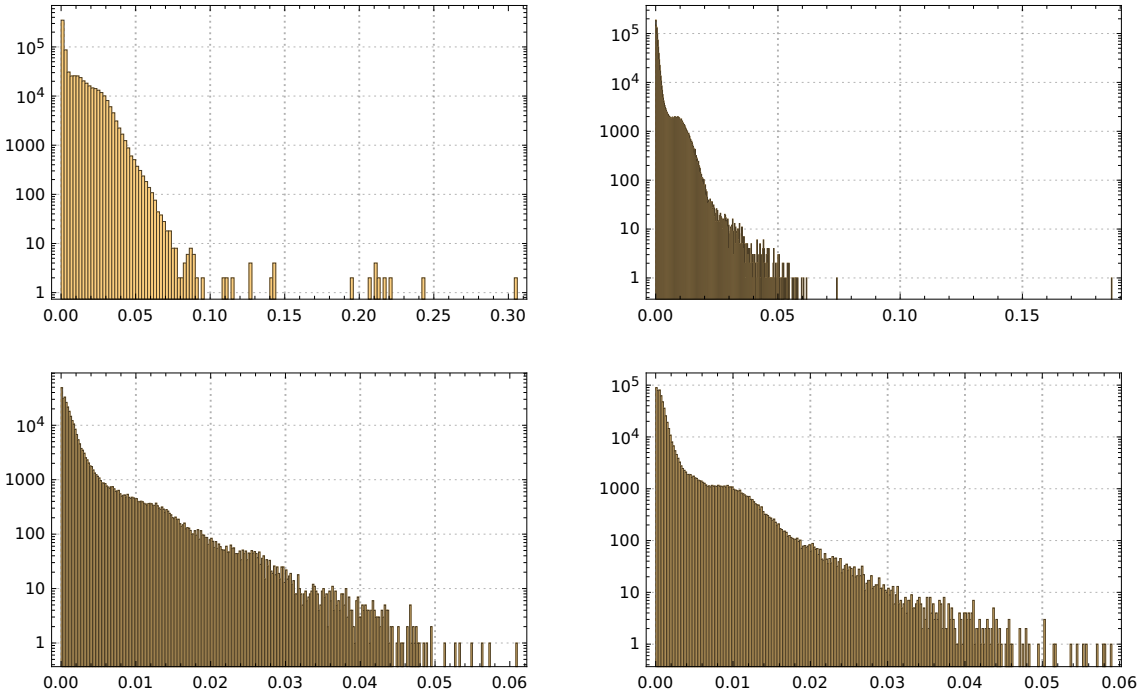
Out[ ]=



## EC vs EC, BS, KP, SA

In[ ]:=
```
GraphicsGrid[ArrayReshape[
  Histogram[#, ScalingFunctions → "Log", PlotTheme → "Detailed", PlotRange → Full] & /@ {
    ecSimilarities,
    Flatten[distanceMat[[ecIndices, bsIndices]]],
    Flatten[distanceMat[[ecIndices, kpIndices]]],
    Flatten[distanceMat[[ecIndices, saIndices]]]
   }, {2, 2}]]
```

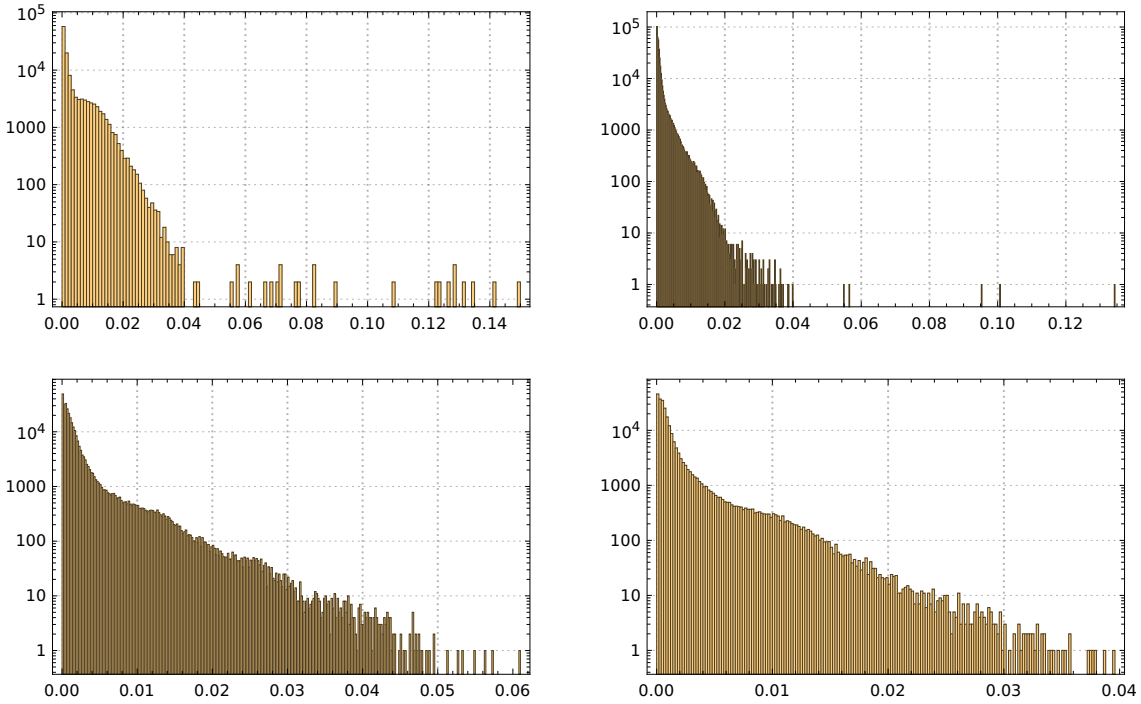
Out[ ]=



## KP vs KP, BS, EC, SA

In[ ]:=
```
GraphicsGrid[ArrayReshape[
  Histogram[#, ScalingFunctions → "Log", PlotTheme → "Detailed", PlotRange → Full] & /@ {
    kpSimilarities,
    Flatten[distanceMat[[kpIndices, bsIndices]]],
    Flatten[distanceMat[[kpIndices, ecIndices]]],
    Flatten[distanceMat[[kpIndices, saIndices]]]
   }, {2, 2}]]
```

Out[ ]=



## SA vs SA, BS, EC, KP

In[ ]:=
```
GraphicsGrid[ArrayReshape[
  Histogram[#, ScalingFunctions → "Log", PlotTheme → "Detailed", PlotRange → Full] & /@ {
    saSimilarities,
    Flatten[distanceMat[[saIndices, bsIndices]]],
    Flatten[distanceMat[[saIndices, ecIndices]]],
    Flatten[distanceMat[[saIndices, kpIndices]]]
   }, {2, 2}]]
```

Out[ ]=