SUP MAT: Evolution of cross-tolerance to metals in yeast

Anna L.Bazzicalupo, Penelope C. Kahn, Eully Ao, Joel Campbell, Sarah P.Otto

Calculating the chance of hitting the same gene once or twice

Given the 6607 genes in the yeast genome, the chance that all genic mutations observed in a metal occur in different genes:

In[45]:=
```
pr1[mutations_] := Product[(6607 - (i - 1))/6607, {i, 1, mutations}] // N
```

The chance that one gene is hit twice (but all other mutations occur in unique genes):

In[46]:=
```
pr2[mutations_] := Total[Table[Product[(6607 - (i - 1))/6607, {i, 1, j - 1}] * ((j - 1)/6607) *
    Product[(6607 - (i - 2))/6607, {i, j + 1, mutations}] // N, {j, 2, mutations}]]
```

The first mutation occurs anywhere. This sums together the chance that the second, third, fourth etc. mutation occurs in one of the previous j-1 genes. Even though the jth gene hits a gene for the second time, all other mutations hit unique genes (contributing to the products).

For the cobalt environment, there are so many mutations (75) that it becomes reasonably likely that at least one gene would be hit twice and so we must consider also the probability that there are two genes hit twice:

In[47]:=
```
pr22[mutations_] := Total[
  Flatten[Table[Product[(6607 - (i - 1))/6607, {i, 1, j - 1}] * ((j - 1)/6607) * Product[(6607 - (i - 2))/6607,
      {i, j + 1, k - 1}] * ((k - 1)/6607) * Product[(6607 - (i - 3))/6607, {i, k + 1, mutations}] // N,
    {j, 2, mutations - 1}, {k, j + 1, mutations}]]]
```

Data

The following table gives the number of times a gene was hit once, twice, etc. for each metal, as well as for the full data set (“all”):

In[48]:=
```
data = {{"#hit", "all", "cd", "co", "cu", "mn", "ni", "zn"},
    {1, 195, 37, 70, 16, 51, 17, 18}, {2, 11, 2, 3, 0, 2, 2, 2}, {3, 2, 1, 0, 0, 1, 1, 1},
    {4, 1, 0, 1, 0, 0, 0, 0}, {5, 3, 0, 0, 0, 0, 0, 1}, {6, 0, 0, 1, 0, 0, 0, 0},
    {7, 0, 0, 0, 0, 0, 0, 0}, {8, 1, 0, 0, 0, 0, 0, 0}, {9, 1, 0, 0, 0, 1, 0, 0}};
MatrixForm[
 data]
```

Out[=]//MatrixForm=

#hit	all	cd	co	cu	mn	ni	zn
1	195	37	70	16	51	17	18
2	11	2	3	0	2	2	2
3	2	1	0	0	1	1	1
4	1	0	1	0	0	0	0
5	3	0	0	0	0	0	1
6	0	0	1	0	0	0	0
7	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0
9	1	0	0	0	1	0	0

Analyses

Cadmium - significant parallelism

In[9]:=
```
data[[All, 3]][[1]]
```

Out[9]=
```
cd
```

In[10]:=
```
Drop[data[[All, 3]], 1]
```

Out[10]=
```
{37, 2, 1, 0, 0, 0, 0, 0, 0}
```

In[11]:=
```
nummut = Total[%]
```

Out[11]=
```
40
```

Of these, the number of multiply hit genes were:

In[12]:=
```
nummultiple = % - %%[[1]]
```

Out[12]=
```
3
```

The probability that all mutations hit separate genes is:

In[13]:=
```
pr1[nummut]
```

Out[13]=
```
0.888436
```

The probability that either all mutations hit separate genes or at most one gene is hit twice is:

In[14]:=
```
pr1[nummut] + pr2[nummut]
```

Out[14]=
```
0.993945
```

Thus, $p<0.01$ for seeing more than one gene hit multiple times (there were three for cadmium), let alone having a gene hit more than two times (one gene hit three times).

Cobalt - significant parallelism [had to include up to two genes hit twice]

In[15]:=
```
data[[All, 4]][[1]]
```
Out[15]=
```
co
```

In[16]:=
```
Drop[data[[All, 4]], 1]
```
Out[16]=
```
{70, 3, 0, 1, 0, 1, 0, 0, 0}
```

In[17]:=
```
nummut = Total[%]
```
Out[17]=
```
75
```

In cobalt, more mutations accumulated.

Of these, the number of multiply hit genes were:

In[18]:=
```
nummultiple = % - %%[[1]]
```
Out[18]=
```
5
```

The probability that all mutations hit separate genes is:

In[19]:=
```
pr1[nummut]
```
Out[19]=
```
0.655999
```

The probability that either all mutations hit separate genes or at most one gene is hit twice is:

In[20]:=
```
pr1[nummut] + pr2[nummut]
```
Out[20]=
```
0.934645
```

Thus, there are so many hits to cobalt that it's still reasonably likely (p>0.01) that we would see one double hit. Allowing two double hits though, the observed data is unlikely to be seen.

In[21]:=
```
pr1[nummut] + pr2[nummut] + pr22[nummut]
```
Out[21]=
```
0.992757
```

That is, p<0.01 for seeing more than two genes hit multiple times (there were five for cobalt), let alone having a gene hit more than two times.

Copper - no parallelism

In[22]:=
```
data[[All, 5]][[1]]
```
Out[22]=
```
cu
```

In[23]:=
```
Drop[data[[All, 5]], 1]
```
Out[23]=
```
{16, 0, 0, 0, 0, 0, 0, 0, 0}
```

In[24]:=
```
nummut = Total[%]
```
Out[24]=
```
16
```

Of these, the number of multiply hit genes were:

In[25]:=
```
nummultiple = % - %%[[1]]
```
Out[25]=
```
0
```

The probability that all mutations hit separate genes is:

In[26]:=
```
pr1[nummut]
```
Out[26]=
```
0.981987
```

Thus, it is very likely (p>0.98) to see no multiply-hit genes (none were observed for copper).

Manganese - significant parallelism

In[28]:=
```
data[[All, 6]][[1]]
```
Out[28]=
```
mn
```

In[29]:=
```
Drop[data[[All, 6]], 1]
```
Out[29]=
```
{51, 2, 1, 0, 0, 0, 0, 0, 1}
```

In[30]:=
```
nummut = Total[%]
```
Out[30]=
```
55
```

Of these, the number of multiply hit genes were:

In[31]:=
```
nummultiple = % - %%[[1]]
```
Out[31]=
```
4
```

The probability that all mutations hit separate genes is:

In[32]:=
```
pr1[nummut]
```
Out[32]=
```
0.798211
```

The probability that either all mutations hit separate genes or at most one gene is hit twice is:

In[33]:=
```
pr1[nummut] + pr2[nummut]
```
Out[33]=
```
0.979097
```

Thus, p<0.03 for seeing more than one gene hit multiple times (there were four for manganese), let alone having genes hit more than two times.

Nickle - significant parallelism

In[34]:=
```
data[[All, 7]][[1]]
```
Out[34]=
```
ni
```

In[35]:=
```
Drop[data[[All, 7]], 1]
```
Out[35]=
```
{17, 2, 1, 0, 0, 0, 0, 0, 0}
```

In[36]:=
```
nummut = Total[%]
```
Out[36]=
```
20
```

Of these, the number of multiply hit genes were:

In[37]:=
```
nummultiple = % - %%[[1]]
```
Out[37]=
```
3
```

The probability that all mutations hit separate genes is:

In[38]:=
```
pr1[nummut]
```
Out[38]=
```
0.971625
```

The probability that either all mutations hit separate genes or at most one gene is hit twice is:

In[39]:=
```
pr1[nummut] + pr2[nummut]
```
Out[39]=
```
0.999647
```

Thus, p<0.001 for seeing more than one gene hit multiple times (there were three for nickel), let alone having genes hit more than two times.

Cadmium - significant parallelism

In[40]:=
```
data[[All, 8]][[1]]
```
Out[40]=
```
zn
```

In[41]:=
```
Drop[data[[All, 8]], 1]
```
Out[41]=
```
{18, 2, 1, 0, 1, 0, 0, 0, 0}
```

In[42]:=
```
nummut = Total[%]
```
Out[42]=
```
22
```

Of these, the number of multiply hit genes were:

In[43]:=
```
nummultiple = % - %%[[1]]
```
Out[43]=
```
4
```

The probability that all mutations hit separate genes is:

In[44]:=
```
pr1[nummut]
```
Out[44]=
```
0.965605
```

The probability that either all mutations hit separate genes or at most one gene is hit twice is:

In[45]:=
```
pr1[nummut] + pr2[nummut]
```
Out[45]=
```
0.999473
```

Thus, p<0.001 for seeing more than one gene hit multiple times (there were four for zinc), let alone having genes hit more than two times.

All metals together - significant parallelism

```
In[•]:= data[[All, 2]][[1]]
Out[•]= all

In[•]:= Drop[data[[All, 2]], 1]
Out[•]= {195, 11, 2, 1, 3, 0, 0, 1, 1}

In[•]:= nummut = Total[%]
Out[•]= 214
```

Of these, the number of multiply hit genes were:

```
In[•]:= nummultiple = % - %%[[1]]
Out[•]= 19
```

The probability that all mutations hit separate genes is:

```
In[•]:= pr1[nummut]
Out[•]= 0.0305836
```

The probability that either all mutations hit separate genes or at most one gene is hit twice is:

```
In[•]:= pr1[nummut] + pr2[nummut]
Out[•]= 0.139597
```

Thus, there are so many hits overall that it's still reasonably likely (p>0.01) that we would see one double hit. Allowing two double hits isn't even enough:

```
In[•]:= pr1[nummut] + pr2[nummut] + pr22[nummut]
Out[•]= 0.332639
```

We thus switch to simulating data draws from a multinomial.

```
In[•]:= SeedRandom[2120]
```

Generating random draws of nummut mutations, each of which could occur in any one of 6607 possible genes (with an equal probability of mutating), repeating this 1000 times:

```
In[•]:= tab = Table[1 / 6607, {i, 1, 6607}];
rantab = Table[BinCounts[
    RandomInteger[MultinomialDistribution[nummut, tab]], {0, 10, 1}], {i, 1, 1000}];
```

The 95% quantile for expected number of double hit genes is 0-7 (median of 3), whereas 11 were observed:

```
In[•]:= {Quantile[rantab[[All, 3]], 0.025],
 Quantile[rantab[[All, 3]], 0.5], Quantile[rantab[[All, 3]], 0.975]}
Out[•]= {0, 3, 7}
```

The 95% quantile for expected number of triple-plus hit genes is 0-1 (median of 0), whereas 8 were observed:

```
In[•]:= {Quantile[Sum[rantab[[All, i]], {i, 4, 10}], 0.025],
         Quantile[Sum[rantab[[All, i]], {i, 4, 10}], 0.5],
         Quantile[Sum[rantab[[All, i]], {i, 4, 10}], 0.975]}
Out[•]= {0, 0, 1}

In[•]:= Max[rantab[[All, 4]]]
Out[•]= 1
```

Only 3% of simulations had any genes hit more than twice, whereas 8 were observed:

```
In[•]:= Total[Sum[rantab[[All, i]], {i, 4, 10}]] / 1000.
Out[•]= 0.03
```