Project I: Determining Mutation Rates

Learning Objectives

The student will be able to

- manipulate sequence data in Excel
- visualize data using a scatterplot
- find the best fit straight line together with the equation and R^2 value
- determine the mutation rate

Knowledge and Skills

- Nucleic acid nomenclature
- Single nucleotide polymorphism (SNP)
- Excel functions: IF, COUNTIF, COUNTBLANK
- Scatterplot
- Best fit straight line

Prerequisites

- Copying and pasting cell content in Excel
- Relative and absolute cell addresses in Excel
- Filling a series in Excel
- Some familiarity with functions and graphing in Excel
- Straight line equation and slope of a straight line

Source: Lieberman, T.D. *et al.* Parallel bacterial evolution within multiple patients identifies candidate pathogenicity genes. *Nature Genetics* 43: 1174-1176 doi:10.1038/ng.1011 (http://www.nature.com/ng/journal/v43/n12/full/ng.997.html)

A retrospective study of 112 isolates from 14 individuals with cystic fibrosis collected during an outbreak of *Burkholderia dolosa* over 16 years identifies genetic mutations.

Task: Determine the number of SNPs between each isolate and the outgroup as a function of the years since the first strain was isolated. Use a linear fit to estimate the number of mutations fixed per year.

Citation: Neuhauser, C. Project I: Determining Mutation Rates.
Created: January 7, 2012 **Revisions:**

Nucleic Acid Nomenclature

http://en.wikipedia.org/wiki/Nucleic_acid_nomenclature

Letter	Nucleotide(s) included
A	A
T	T
G	G
C	C
R	G or A
Y	T or C
M	A or C
K	G or T
S	G or C
W	A or T
H	A or C or T
B	G or T or C
V	G or C or A
D	G or T or A
N	G or T or A or C

Step 1:

Download the Supplementary Table 2 from Nature Genetics:
http://www.nature.com/ng/journal/v43/n12/full/ng.997.html#/supplementary-information

The spreadsheet contains the sequence information of the 112 epidemic *Burkholderia dolosa* isolates from 14 subjects.

Copy the data from the tab Supplementary Table 2 into a new sheet. Use the Paste Values option.

For each sequence determine the number of sites that are not uniquely identified (see letter code above) and the number of sites with insertion or deletions (blanks). The following Excel functions will be useful:

- COUNTIF(*range*, *condition*)
- COUNTBLANK(*range*)

The following table includes the counts for the outgroup and the isolate A-0-0:

Citation: Neuhauser, C. Project I: Determining Mutation Rates.
Created: January 7, 2012 **Revisions:**

	A	B	C
1		Count	
2	Letter	X-3-7 (outgroup)	A-0-0
3	A	95	86
4	T	91	95
5	G	158	163
6	C	167	164
7	R	0	0
8	Y	0	1
9	M	0	0
10	K	0	0
11	S	0	0
12	W	0	0
13	H	0	0
14	B	0	0
15	V	0	0
16	D	0	0
17	N	0	0
18	Blank	0	2

For instance, to find the number of sites with As, we use
Cell B3: COUNTIF(*range*,A3)
To count the number of blanks, we use
Cell B18: COUNTBLANK(*range)*

Determine this information for all the sequences to prepare for counting the single nucleotide polymorphisms.

Single Nucleotide Polymorphism

http://www.ornl.gov/sci/techresources/Human_Genome/faq/snps.shtml

To determine the number of SNPs per genome as a function of time for each of the 112 isolates, we need to count the number of differences between an isolate and a reference genome. We choose the outgroup as the reference genome, and compare each isolate to the aligned outgroup sequence. We count the number of sites where the nucleotide differs between the outgroup and the isolate. We do not count sites with indels or where the nucleotide could not be identified.). The following Excel function will be useful:

- IF(*logical test*, [*value if true*], [*value if false*])

Citation: Neuhauser, C. Project I: Determining Mutation Rates.
Created: January 7, 2012 **Revisions:**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	X-3-7 (outgroup)	A	C	G	C	G	C	T	C	T	T	A	T	G	T	G
2	A-0-0	A	G	G	C	G	C	T			T	A	T	G	Y	G
3	Difference Y/N	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0

To count the number of differences, excluding indels, we use the IF function. For instance, assuming that Cell Q1 is blank,

Cell B3: IF(B1=B2,0,1)*IF(B1=Q1,0,1)*IF(B2=Q1,0,1)
Cell C3: IF(C1=C2,0,1)*IF(C1=Q1,0,1)*IF(C2=Q1,0,1)

In order to drag cells, we fix rows and/or columns as appropriate, using the $ operator:

Cell B3: IF(B$1=B2,0,1)*IF(B$1=$Q$1,0,1)*IF(B2=$Q$1,0,1)
We can then drag the cell B3 across the sites of all isolates and fill the spreadsheet with 0s and 1s to indicate potential SNPs.

Since the isolates have sites with letters other than the four nucleotides, we need to subtract those sites from the count.

Suppose now that the number of differences excluding indels between the outgroup and the isolate A-0-0 is 32 and the number of sites with letters other than the four nucleotides is 1, then the number of SNPs is 31.

Step 2:
Count the number of sites where the two sequences differ (excluding indels) and subtract the number of sites where the nucleotide was not identified using the functions from Step 1.

Step 3:
Graph the number of SNPs as a function of time and find the mutation rate.

Hint:
Find the number of SNPs between the outgroup and each of the isolates. Transfer the counts to a new sheet, and calculate the number of years since the first collection (patient 0, isolate A-0-0). Plot the counts as a function of time in a scatterplot, and find the equation of the best fit straight line together with the R^2 value. The slope of the line is related to the mutation rate (why?).

Excel Hints

Your data for a scatterplot must be in the form (x, y), one column for the x values, another column for the y values. Select the cells that contain the data. Click the **Insert** tab, choose the **Scatter** option in the **Charts** group. Select the scatter plot with only markers.

Citation: Neuhauser, C. Project I: Determining Mutation Rates.
Created: January 7, 2012 **Revisions:**

If you click on the graph, the **Chart Tools** tabs appear. Click on the **Layout** tab. To fit a trendline, click on **Trendline** in the **Analysis group**. Select **More Trendline Options…** and select the Linear regression chart. Select the Display Equation on chart on Display R-squared value on chart.

Further Study

Step 4:

Consult the literature to determine whether the rates are consistent with other human infections.

Morelli, G. *et al.* Microevolution of *Heliobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet*. **6**, e1001036 (2010).

Smith, E.E. *et al.* Genetic adaptation by *Pseudomonas aeruginosa* to the airways of cystic fibrosis patients. *Proc. Natl. Acad. Sci. USA* **103**, 8487-8492 (2006).

Citation: Neuhauser, C. Project I: Determining Mutation Rates.
Created: January 7, 2012 **Revisions:**