Bacterial Evolution

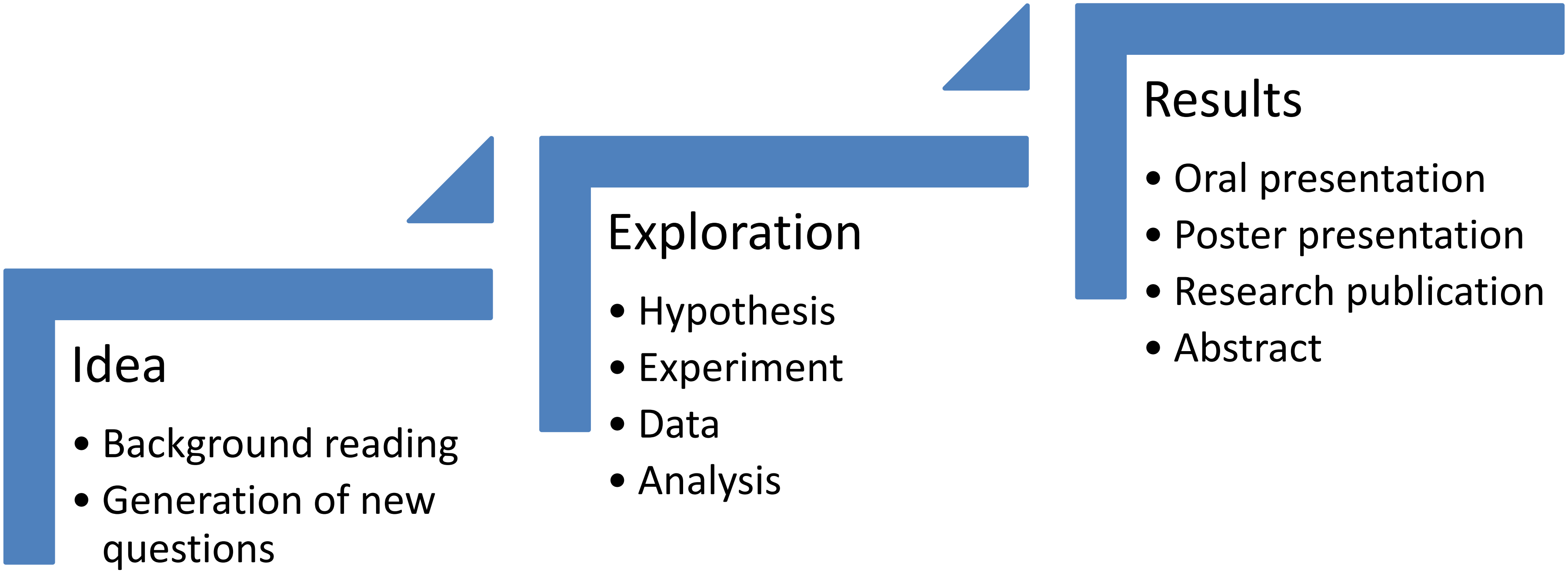
Turning a scientific paper into an open-ended classroom experience

Claudia Neuhauser

University of Minnesota Rochester

January 2012

Components of Research



Idea

- Background reading
- Generation of new questions

Exploration

- Hypothesis
- Experiment
- Data
- Analysis

Results

- Oral presentation
- Poster presentation
- Research publication
- Abstract

Developing Research Skills Backwards

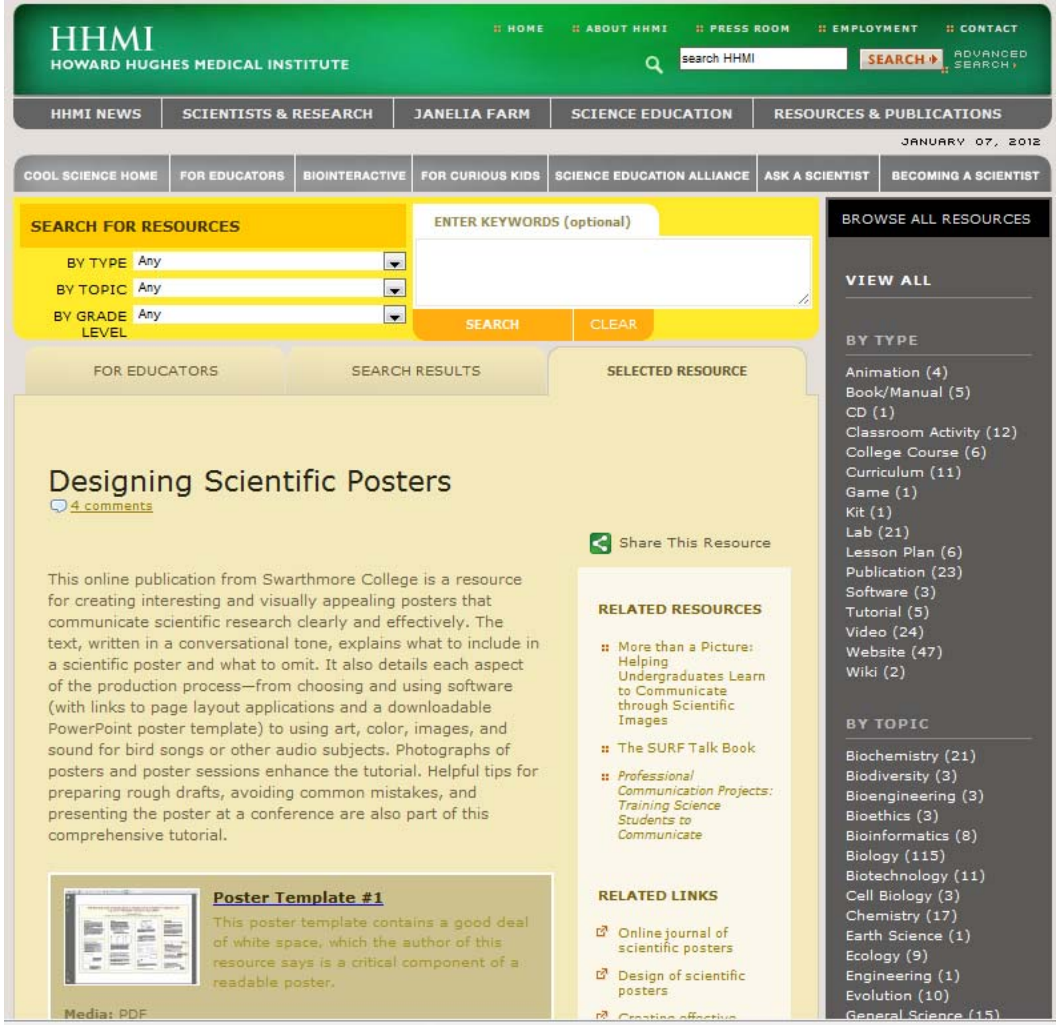
- Reading a scientific paper
  - Focus on a key result and
    - produce a poster
    - give a short oral presentation
    - write an abstract
    - turn into a proposal
    - write a popular science piece
- Synthesizing multiple scientific papers
  - Focus on a key insight and
    - identify relevant resources
    - write a review paper

Developing Research Skills Backwards

- Reading a scientific paper
  - Focus on a key result and
    - **produce a poster**
    - give a short oral presentation
    - **write an abstract**
    - turn into a proposal
    - write a popular science piece
- Synthesizing multiple scientific papers
  - Focus on a key insight and
    - identify relevant resources
    - write a review paper

…produce a poster

- Paper-scissors-glue
  - Cut up a paper, paste the pieces on a poster template
- HHMI Cool Science
- http://www.hhmi.org/coolscience/resources/SPT--FullRecord.php?ResourceId=29



JANUARY 07, 2012

BY TYPE: Any
BY TOPIC: Any
BY GRADE LEVEL: Any

Designing Scientific Posters

4 comments

This online publication from Swarthmore College is a resource for creating interesting and visually appealing posters that communicate scientific research clearly and effectively. The text, written in a conversational tone, explains what to include in a scientific poster and what to omit. It also details each aspect of the production process—from choosing and using software (with links to page layout applications and a downloadable PowerPoint poster template) to using art, color, images, and sound for bird songs or other audio subjects. Photographs of posters and poster sessions enhance the tutorial. Helpful tips for preparing rough drafts, avoiding common mistakes, and presenting the poster at a conference are also part of this comprehensive tutorial.

Poster Template #1

This poster template contains a good deal of white space, which the author of this resource says is a critical component of a readable poster.

Media: PDF

RELATED RESOURCES

- More than a Picture: Helping Undergraduates Learn to Communicate through Scientific Images
- The SURF Talk Book
- *Professional Communication Projects: Training Science Students to Communicate*

RELATED LINKS

- Online journal of scientific posters
- Design of scientific posters
- Creating effective

BY TYPE

Animation (4)
Book/Manual (5)
CD (1)
Classroom Activity (12)
College Course (6)
Curriculum (11)
Game (1)
Kit (1)
Lab (21)
Lesson Plan (6)
Publication (23)
Software (3)
Tutorial (5)
Video (24)
Website (47)
Wiki (2)

BY TOPIC

Biochemistry (21)
Biodiversity (3)
Bioengineering (3)
Bioethics (3)
Bioinformatics (8)
Biology (115)
Biotechnology (11)
Cell Biology (3)
Chemistry (17)
Earth Science (1)
Ecology (9)
Engineering (1)
Evolution (10)
General Science (15)

…write an abstract

1. Topic
2. Purpose
3. Method
4. Results
5. Discussion and conclusion

Bacterial pathogens evolve during the infection of their human host[1, 2, 3, 4, 5, 6,7, 8], but separating adaptive and neutral mutations remains challenging[9, 10, 11]. Here we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple individuals. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among subjects with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired nonsynonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes affect important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition and implicate oxygen-dependent regulation as paramount in lung infections. Several genes have not previously been implicated in pathogenesis and may represent new therapeutic targets. The identification of parallel molecular evolution as a pathogen spreads among multiple individuals points to the key selection forces it experiences within human hosts.

…write an abstract

1. Topic
2. Purpose
3. Method
4. Results
5. Discussion and conclusion

Bacterial pathogens evolve during the infection of their human host[1, 2, 3, 4, 5, 6,7, 8], but separating adaptive and neutral mutations remains challenging[9, 10, 11]. Here we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple individuals. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among subjects with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired nonsynonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes affect important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition and implicate oxygen-dependent regulation as paramount in lung infections. Several genes have not previously been implicated in pathogenesis and may represent new therapeutic targets. The identification of parallel molecular evolution as a pathogen spreads among multiple individuals points to the key selection forces it experiences within human hosts.

…write an abstract

1. Topic
2. Purpose
3. Method
4. Results
5. Discussion and conclusion

Bacterial pathogens evolve during the infection of their human host[1, 2, 3, 4, 5, 6,7, 8], but separating adaptive and neutral mutations remains challenging[9, 10, 11]. Here we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple individuals. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among subjects with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired nonsynonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes affect important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition and implicate oxygen-dependent regulation as paramount in lung infections. Several genes have not previously been implicated in pathogenesis and may represent new therapeutic targets. The identification of parallel molecular evolution as a pathogen spreads among multiple individuals points to the key selection forces it experiences within human hosts.

…write an abstract

1. Topic
2. Purpose
3. Method
4. Results
5. Discussion and conclusion

Bacterial pathogens evolve during the infection of their human host[1, 2, 3, 4, 5, 6, 7, 8], but separating adaptive and neutral mutations remains challenging[9, 10, 11]. Here we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple individuals. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among subjects with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired nonsynonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes affect important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition and implicate oxygen-dependent regulation as paramount in lung infections. Several genes have not previously been implicated in pathogenesis and may represent new therapeutic targets. The identification of parallel molecular evolution as a pathogen spreads among multiple individuals points to the key selection forces it experiences within human hosts.

…write an abstract

1. Topic
2. Purpose
3. Method
4. Results
5. Discussion and conclusion

Bacterial pathogens evolve during the infection of their human host[1, 2, 3, 4, 5, 6,7, 8], but separating adaptive and neutral mutations remains challenging[9, 10, 11]. Here we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple individuals. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among subjects with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired nonsynonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes affect important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition and implicate oxygen-dependent regulation as paramount in lung infections. Several genes have not previously been implicated in pathogenesis and may represent new therapeutic targets. The identification of parallel molecular evolution as a pathogen spreads among multiple individuals points to the key selection forces it experiences within human hosts.

…write an abstract

1. Topic
2. Purpose
3. Method
4. Results
5. Discussion and conclusion

Bacterial pathogens evolve during the infection of their human host[1, 2, 3, 4, 5, 6,7, 8], but separating adaptive and neutral mutations remains challenging[9, 10, 11]. Here we identify bacterial genes under adaptive evolution by tracking recurrent patterns of mutations in the same pathogenic strain during the infection of multiple individuals. We conducted a retrospective study of a *Burkholderia dolosa* outbreak among subjects with cystic fibrosis, sequencing the genomes of 112 isolates collected from 14 individuals over 16 years. We find that 17 bacterial genes acquired nonsynonymous mutations in multiple individuals, which indicates parallel adaptive evolution. Mutations in these genes affect important pathogenic phenotypes, including antibiotic resistance and bacterial membrane composition and implicate oxygen-dependent regulation as paramount in lung infections. Several genes have not previously been implicated in pathogenesis and may represent new therapeutic targets. The identification of parallel molecular evolution as a pathogen spreads among multiple individuals points to the key selection forces it experiences within human hosts.

RESEARCH IN THE CLASSROOM

Research in the Classroom: Hypothesis→Experiment→Data→Analysis

- Research in a resource-poor environment
  - Experiments are expensive
  - What can we do without experiments?

Research in the Classroom: Hypothesis→Experiment→Data→Analysis

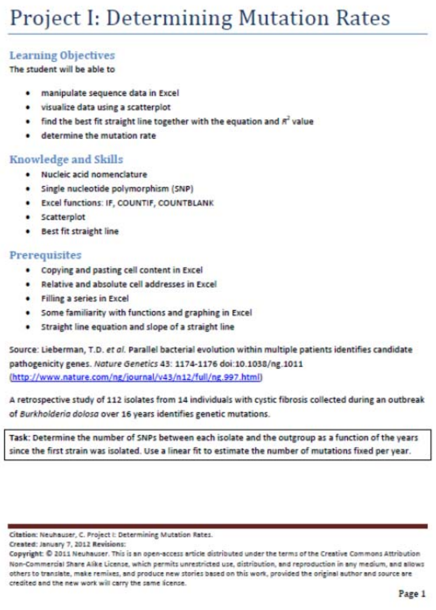
- (Hypothesis→Experiment)→Data→Analysis
  - Assign scientific paper as background reading
  - Provide supplementary data of a scientific paper for re-analysis
- Hypothesis→(Experiment)→Data→ Analysis
  - Introduce students to topic
  - Students develop hypothesis and find publicly available data
  - Students find way to analyze data
- Data→Hypothesis→(Experiment)→Analysis
  - Provide data and context
  - Students develop questions that can be answered with the data
  - Students develop hypothesis
  - Students find way to analyze data

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

PROJECT I

The Worksheet



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:**

- Open-access article distributed under the Creative Commons Non-Commercial Share Alike License

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

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

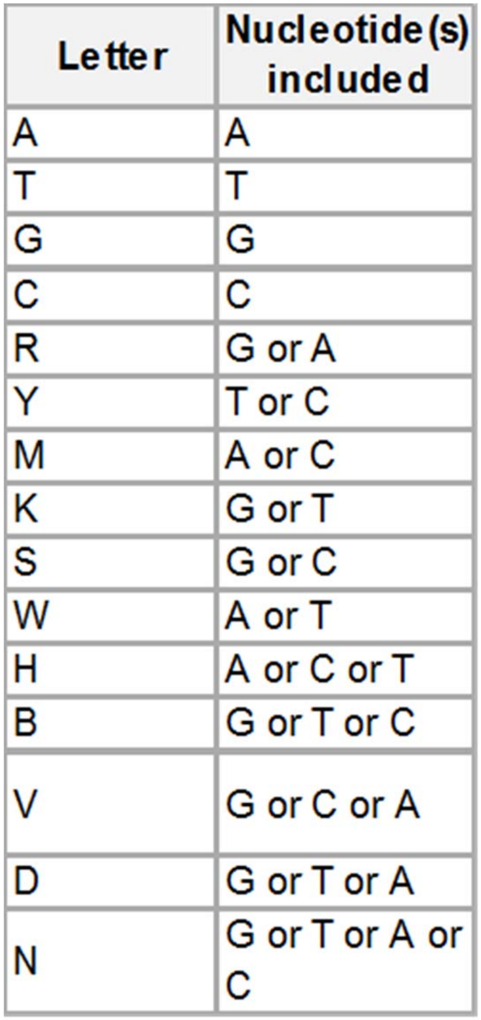
Project I

- Classroom Experiences
  - (Hypothesis→Experiment)→Data→Analysis
    - Assign scientific paper as background reading
    - Provide supplementary data of a scientific paper for re-analysis
- 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.

Tasks

- **Determine the number of SNPs between each isolate and the outgroup as a function of the years since the first strain was isolated.**
  - Excel skills
- **Use a linear fit to estimate the number of mutations fixed per year.**
  - Linear regression

Nucleic Acid Nomenclature



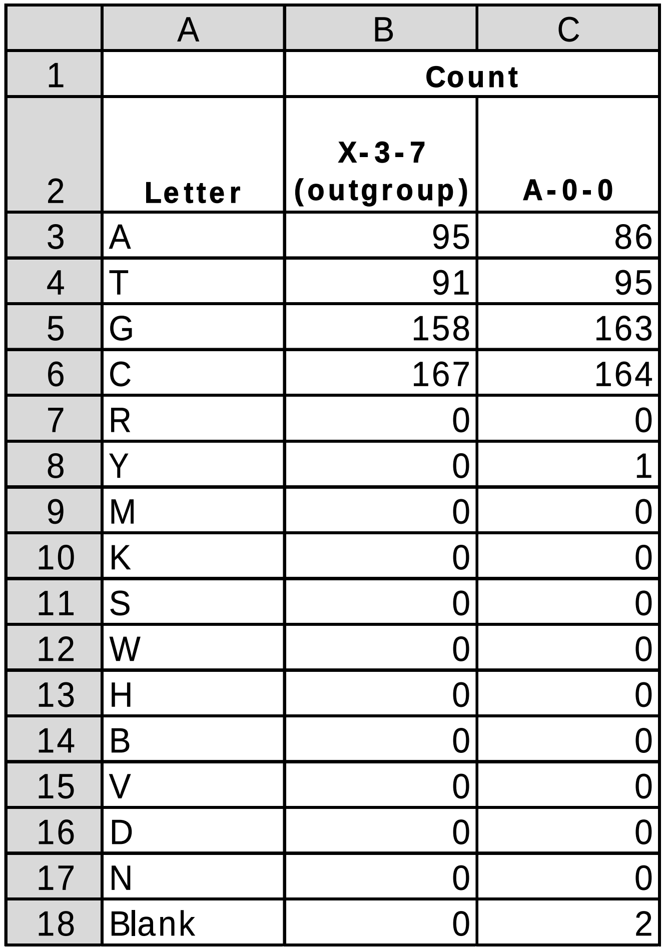
Letter	Nucleotide(s) included
A	A
T	T
G	G
C	C
R	G or A
Y	T or C
M	A or C
K	G or T
S	G or C
W	A or T
H	A or C or T
B	G or T or C
V	G or C or A
D	G or T or A
N	G or T or A or C

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

Step 1

- Download the Supplementary Table 2 from Nature Genetics: http://www.nature.com/ng/journal/v43/n12/full/ng.997.html#/supplementary-information
- Copy the data from the tab Supplementary Table 2 into a new sheet. Use the Paste Values option.
- Count the letters in each of the sequences

Useful Excel Function

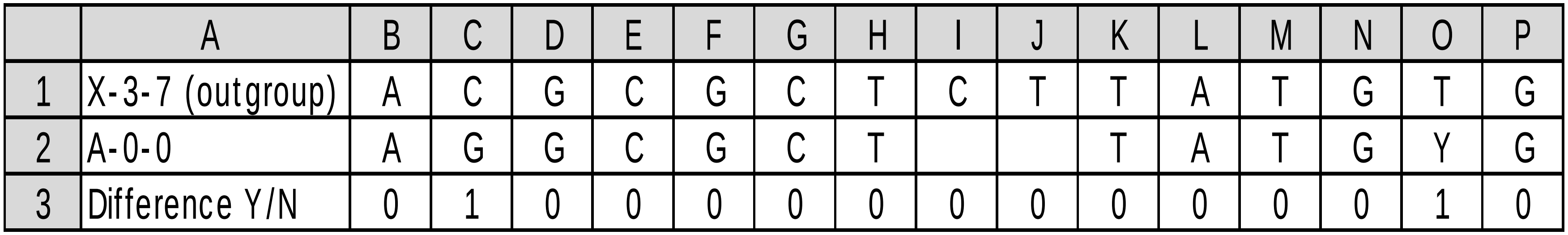


	A	B	C
1		**Count**	
2	**Letter**	**X-3-7 (outgroup)**	**A-0-0**
3	A	95	86
4	T	91	95
5	G	158	163
6	C	167	164
7	R	0	0
8	Y	0	1
9	M	0	0
10	K	0	0
11	S	0	0
12	W	0	0
13	H	0	0
14	B	0	0
15	V	0	0
16	D	0	0
17	N	0	0
18	Blank	0	2

- Cell B3: COUNTIF(*range*,A3)
- To count the number of blanks, we use
- Cell B18: COUNTBLANK(*range)*

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 and .

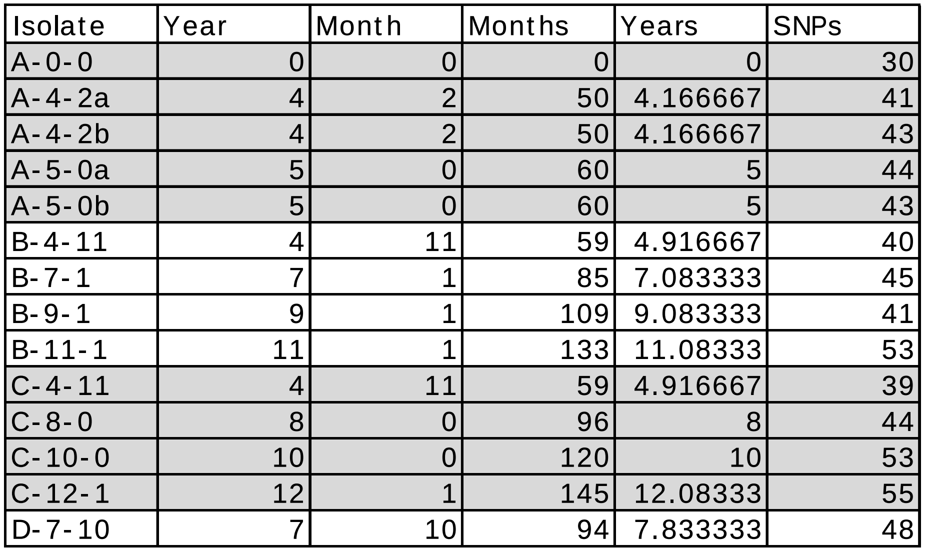


	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	X-3-7 (outgroup)	A	C	G	C	G	C	T	C	T	T	A	T	G	T	G
2	A-0-0	A	G	G	C	G	C	T			T	A	T	G	Y	G
3	Difference Y/N	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0

- To count the number of differences, excluding indels, we use the IF function. For instance, assuming that Cell Q1 is blank (Q1 is an absolute cell),
  - Cell B3: IF(B1=B2,0,1)*IF(B1=$Q$1,0,1)*IF(B2=$Q$1,0,1)
  - Cell C3: IF(C1=C2,0,1)*IF(C1=$Q$1,0,1)*IF(C2=$Q$1,0,1)

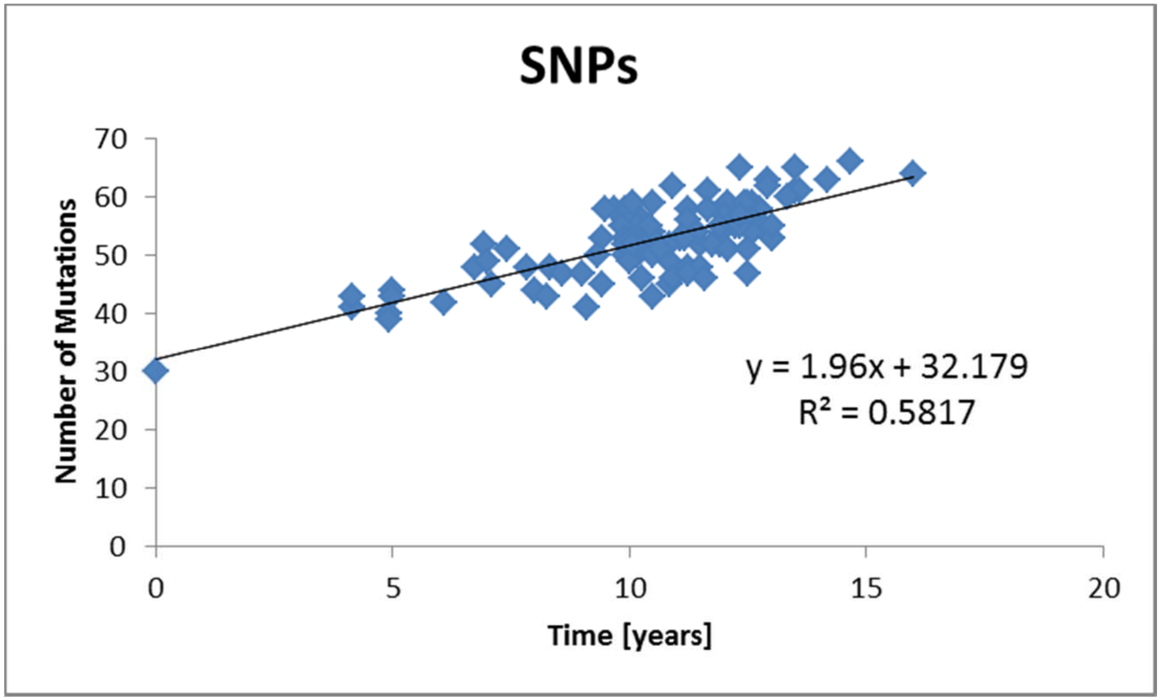
Step 3

- Generate a table that lists the number of SNPs as a function of time.
- Graph the number of SNPs as a function of time and use the best fit straight line to find the mutation rate.



Isolate	Year	Month	Months	Years	SNPs
A-0-0	0	0	0	0	30
A-4-2a	4	2	50	4.166667	41
A-4-2b	4	2	50	4.166667	43
A-5-0a	5	0	60	5	44
A-5-0b	5	0	60	5	43
B-4-11	4	11	59	4.916667	40
B-7-1	7	1	85	7.083333	45
B-9-1	9	1	109	9.083333	41
B-11-1	11	1	133	11.08333	53
C-4-11	4	11	59	4.916667	39
C-8-0	8	0	96	8	44
C-10-0	10	0	120	10	53
C-12-1	12	1	145	12.08333	55
D-7-10	7	10	94	7.833333	48

Graph



- What is the mutation rate?

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

PROJECT II

Bacterial Phylogeny

- **Learning Objectives**
  - The student will be able to
  - convert sequence data from Excel to FASTA file format (optional)
  - align multiple DNA sequences using a general purpose multiple sequence alignment program for DNA sequences (ClustalW2)
  - interpret a phylogenetic tree
- **Knowledge and Skills**
  - FASTA format
  - Online resources for multiple sequence alignment and tree building
  - Phylogenetic trees
- **Prerequisites**
  - Optional: Familiarity with Matlab

Project II

- Classroom Experiences
  - Data→Hypothesis→(Experiment)→Analysis
    - Provide data and context
    - Students develop questions that can be answered with the data
    - Students develop hypothesis
    - Students find way to analyze data
- 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.

Tasks

- **Infer the likely transmission network between individuals based on the bacterial phylogeny of the 112 isolates.**
  - Using online tools to build a phylogeny
  - Interpreting the phylogeny
- **Infer the likely transmission network between lung and blood isolates within individuals based on the bacterial phylogeny of the isolates of individuals K, N, and H.**
  - Using online tools to build a phylogeny
  - Interpreting the phylogeny

FASTA Format

- The FASTA format is a text-based format for DNA or protein sequences that can be read by many software packages.
- Each sequence begins with a single-line starting with the “>” symbol to mark the start of a new sequence, followed by the description of the sequence, and, in a new line, the sequence data.

```
>A-0-0
CGCAATGTCGTCTTCTCCGGCTCTTTGGGCGGCGCCTACTTCG
ACAGCGGGTGGCCTCCTCTGACGACGTCCCGAGCGCAGAAG
```

Source: http://en.wikipedia.org/wiki/FASTA_format

The Data

- The FASTA formatted sequence data is stored in the text file **burkisolates.txt**.
- The first three sequences are the reference sequence AU0158, the outgroup X-3-7, and the Last common ancestor.
- The remaining sequences are the 112 isolates.

Building the Tree: ClustalW2

- Copy the **burkisolates.txt** into a new text file. Call the file **burkphylo.txt**.
- Delete the sequence data of the reference sequence and the last common ancestor.
- We will use ClustalW2 to build the tree. ClustalW2 aligns multiple sequences.
- Go to http://www.ebi.ac.uk/Tools/msa/clustalw2/ and follow the steps. Make sure you select DNA in Step 1. Use the default options in Step 2 and Step 3. Submit your job.
- When your job is ready, you will see the alignment in the same browser window. The Guide Tree tab on top shows the tree (you need to scroll down until the Phylogram).

Tasks

- **Task 1:**
  - Investigate the sequence of transmission among the individuals. Which of the isolates cluster according to subject?
- **Task 2:**
  - Blood is usually a sterile medium, that is, bacteria are not found in the blood. Bacteremia is a condition that indicates bacteria in the blood. Since subjects K, N, and H have isolates from both the airways and the bloodstream, they suffered from bacteremia. In which individuals can you find evidence of multiple transmissions from the lungs to the bloodstream?

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

PROJECT III

Neutral vs. Adaptive Evolution

- **Learning Objectives**
  - The student will be able to
  - build a model for the null hypothesis of a statistical test
- **Knowledge and Skills**
  - Hypothesis testing
  - Neutral versus adaptive evolution
- **Prerequisites**
  - Familiarity with functions and macros in Excel

Project III

- Classroom Experience
  - Hypothesis→(Experiment)→Data→ Analysis
    - Introduce students to topic
    - Students develop hypothesis and find publicly available data
    - Students find way to analyze data
- 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.

Tasks

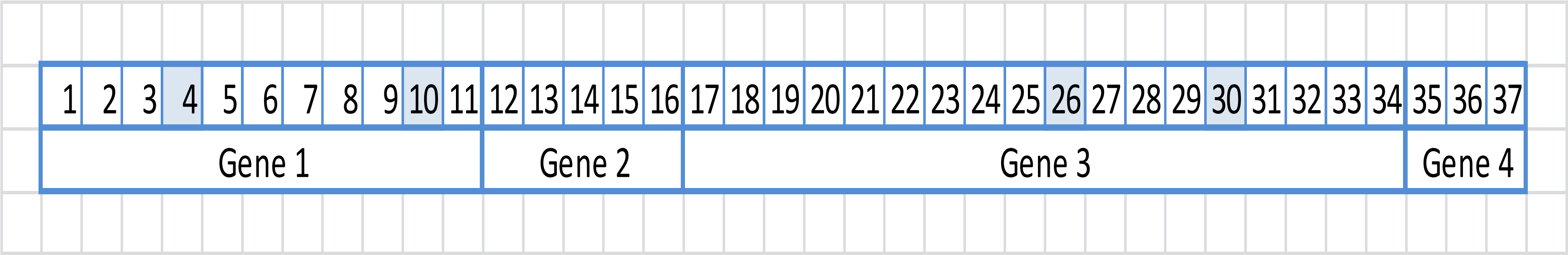
- Build a model to decide whether the observed mutations are consistent with neutral evolution.
  - Formulating a null hypothesis
  - Building a model for the null hypothesis
  - Simulating the model to generate data under the null hypothesis

The Data

- The *B*. *dolosa* genome has **5,014 genes**. Lieberman et al. (2011) observed **561 independent mutational events in 304 genes**.
  - They found that **seventeen genes had three or more different mutations**, and four genes had more than ten different mutations.
- The length of each gene is in the first column of the spreadsheet **burksim.xlsm** (tab Simulation), starting with Cell A3.
  - This spreadsheet is macro enabled, which is a feature that we will need later on when we run simulations.

The Model

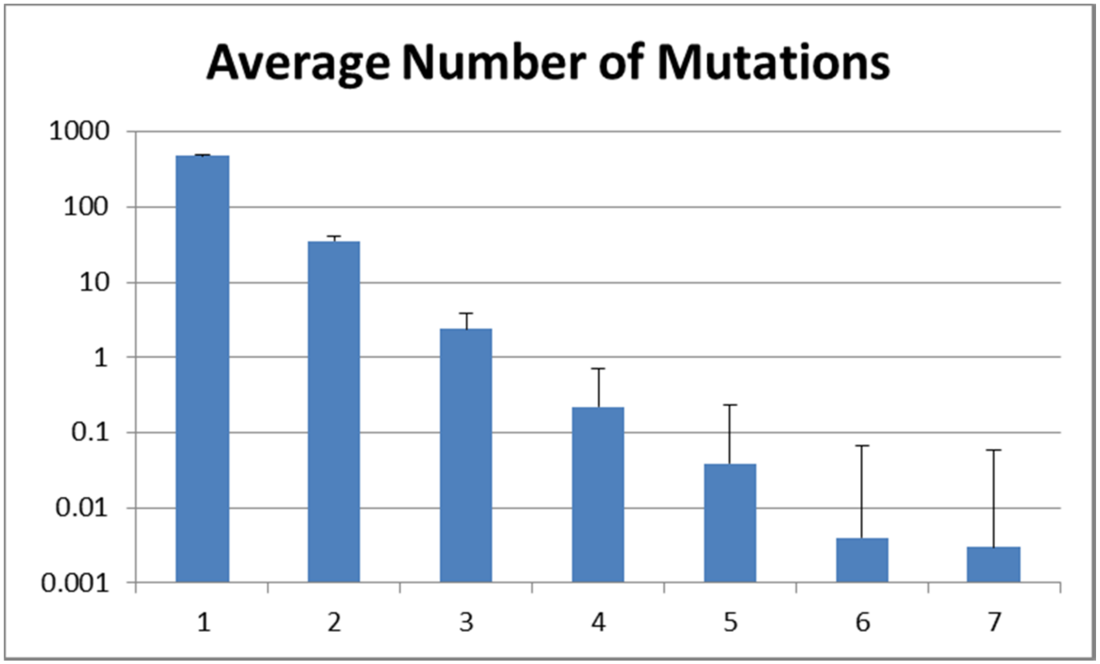
- We think of the genome as a sequence of nucleotides, labeled by location from 1 to the length of the genome.
- We assume that each nucleotide has the same likelihood of being mutated.
  - We simulate the 561 mutations by randomly choosing 561 sites and keeping track which of the 5,014 genes are mutated and how often each one is mutated.



1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37
Gene 1											Gene 2					Gene 3																		Gene 4		

Matlab Simulation

	mean	std
1	482.739	11.2504
2	34.956	5.407708
3	2.407	1.500534
4	0.222	0.470039
5	0.039	0.193692
6	0.004	0.063151
7	0.003	0.054717


Average Number of Mutations