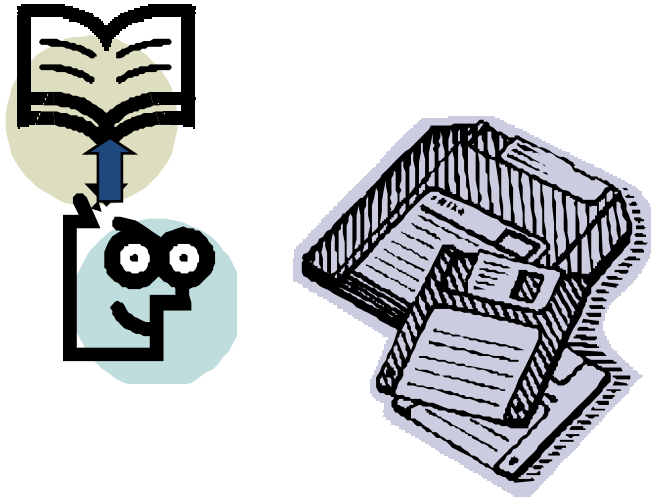
Data Analysis

How to organize and present *your* data into figures and tables

Do you have enough data for a paper?

- Start with organizing your research data into figures and tables.
- Most publications have 5-7 figures and possibly more online.



Data presentation in different settings

1. Presenting to your research group

   The figures/data/findings presented to lab co-workers can be unfinished or raw data

2. Poster presentation

   Depending on formality can include data from other people.

3. Conference oral presentation

   More formal usually with data in easy to view format.

4. Manuscript or submitted paper

   Only original data in figures.

Deciding what data to use

- Ethically what can you do when some samples have problems?
- Issues of technique versus outlying values versus inconsistent results.
- When were samples collected?
- When were samples processed?

What is the most effective way to display your data?

- What type of data do you have?
  - Table or figure or picture
    - » Results should be presented so they are easy to understand and support your main point.
  - A lot of data points?
    - » Results from 5000 tests must be summarized in a way that is easy to understand.

- Try to use a variety in the types of figures.
  - Boring to have 6 identical type of graphs.

Tables or graphs?

Table 2. Blood glucose levels

Time (hour)	Normal (mg/dl*)	Diabetic (mg/dl)
midnight	100.3	175.8
2:00	93.6	165.7
4:00	88.2	159.4
6:00	100.5	72.1
8:00	138.6	271.0
10:00	102.4	224.6
noon	93.8	161.8
2:00	132.3	242.7
4:00	103.8	219.4
6:00	93.6	152.6
8:00	127.8	227.1
10:00	109.2	221.3

* decaliters/milligram

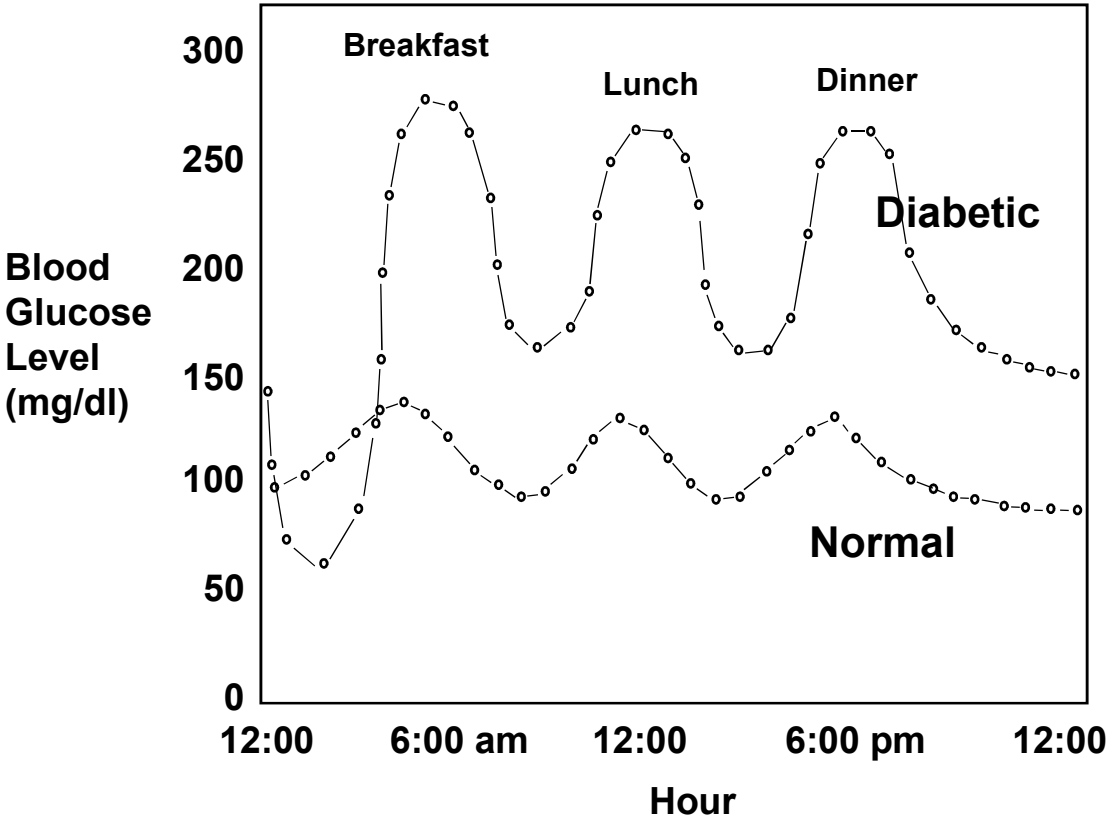


Figure 11. Blood glucose levels over time for normal individual and diabetic subjects.

Graphs are often easier to understand

Table 4. Gastritis Scores of the 175 C. pyloridis Culture-Positive Patients Before and After Therapeutic Regimens

		Treatment									
		CBS		Amoxicillin		CBS + amoxicillin		Cimetidine		Sucralfate	
	Culture	m[a]	n[b]	m	n	m	n	m	n	m	n
Before treatment	+	5.9	67	5.5	22	6.2	20	5.5	53	5.8	13
After treatment	+	3.8	37	2.9	7	6.5	2	5.4	52	5.3	13
	−	1.2[d]	30	0.9	15	1.1[c]	18				
1 mo after treatment	+	5.6	55	4.2	17	5.0	12	5.8	51	5.3	13
	−	0.7[d]	12	0.7[d]	5	0.5[d]	8				
3 mo after treatment	+	5.8	54	5.1	16	5.3	12	5.5	52	5.7	13
	−	0.5[d]	10	0.2[d]	5	0.9[d]	8				
6 mo after treatment	+	5.7	54	5.2	16	6.0	12	5.5	48	5.6	13
	−	0[d]	10	0.6[d]	5	0.6[d]	8				
12 mo after treatment	+	6.6	52	5.5	16	5.5	13	5.3	51	5.9	13
	−	0[d]	10	0[d]	5	0[d]	7				

Gastritis was assessed by scoring for four characteristic pathological parameters of chronic active gastritis (see Methods). +, positive; −, negative. [a] Mean score. [b] Number of patients. [c] $p < 0.05$ vs. culture-positive patients by Wilcoxon rank-sum test. [d] $p < 0.01$ vs. culture-positive patients by Wilcoxon rank-sum test.

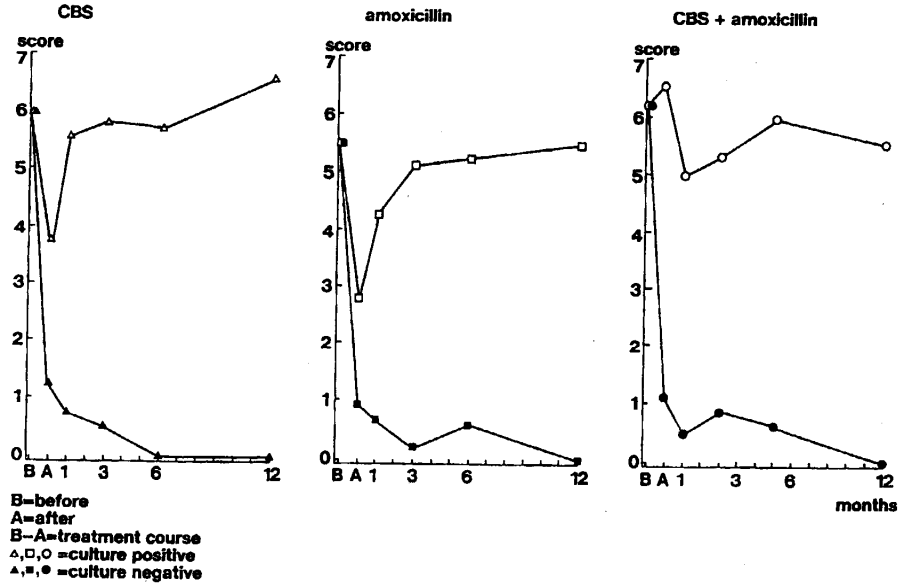


Figure 2. Mean gastritis scores in patients with chronic active gastritis and positive culture for *C. pyloridis*, before (B) and after (A) treatment with CBS, amoxicillin, or a combination of CBS and amoxicillin. Note the ultimate complete disappearance of inflammatory changes of gastric mucosa (gastritis score = 0) after persistent eradication of *C. pyloridis* during follow-up of 1 yr. (See Table 4 for actual values.)

FIGURE 3.9 Two different representations of a data set. Table and figures from Rauws et al. (1988).

Figure Example

How to display the data?

Raw Data from Leishmania paper

A bit confusing in a table.

How to present?

Pt #	Case	IFN-g SLA	IFN-g PHA	IL-2 SLA	IL-2 PHA	IL-10 SLA	IL-10 PHA	IgE
		pg/ml	pg/ml	pg/ml	pg/ml	pg/ml	pg/ml	ng/ml
1	Active ACL	1140	>2000	<125	700	50	>2000	600
2	Active ACL	<125	870	<125	120	<125	208	40
7	Active ACL	<125	755	<125	<100	<125	600	190
8	Active ACL	629	<125	<125	600	140	1160	220
12	Active ACL	>2000	>2000	<125	1100	<125	>2000	140
22	Active ACL	<125	>2000	<125	720	<125	930	40
23	Active ACL	<125	250	<125	<100	80	<125	75
24	Active ACL	<125	1045	<125	200	<125	154	140
25	Active ACL	<125	<125	<125	<100	<125	100	320
26	Active ACL	162	697	<125	<100	<125	468	250
27	Active ACL	<125	426	<125	<100	250	210	226
28	Active ACL	<125	<125	<125	<100	250	225	367
29	Active ACL	<125	777	<125	400	460	260	480
30	Active ACL	<125	313	<125	<100	320	770	134
31	Active ACL	<125	1018	<125	300	250	240	57
36	Active ACL	882	720	411	1093	217	>2000	432
37	Active ACL	1568	>2000	600	741	150	482	341
5	Asymptomatic	<125	<125	<125	<100	<125	340	210
6	Asymptomatic	320	196	<125	<100	<125	690	380
9	Asymptomatic	960	>2000	<125	250	<125	720	150
15	Asymptomatic	460	>2000	<125	1500	<125	>2000	100
18	Asymptomatic	210	>2000	<125	700	<125	1450	<15
19	Asymptomatic	<125	>2000	<125	350	<125	>2000	90
20	Asymptomatic	1440	>2000	<125	1000	<125	>2000	140
21	Asymptomatic	500	>2000	<125	1850	<125	>2000	65
34	Asymptomatic	488	>2000	<125	<125	<125	>2000	615
3	Neg ctrl	<125	>2000	<125	<100	<125	538	480
4	Neg ctrl	<125	>2000	<125	<100	<125	300	330
10	Neg ctrl	<125	106	<125	<100	<125	235	140
11	Neg ctrl	<125	>2000	<125	>2000	130	>2000	250
13	Neg Ctrl	980	>2000	<125	550	<125	>2000	40
14	Neg ctrl	<125	>2000	<125	>2000	<125	>2000	90
16	Neg ctrl	<125	>2000	<125	1800	<125	>2000	180
17	Neg ctrl	220	>2000	<125	>2000	<125	>2000	100
32	Neg Ctrl	553	1709	<125	<125	<125	<125	510
33	Neg Ctrl	337	1909	<125	<125	<125	>2000	554
35	Neg ctrl	<125	300	<125	294	<125	>2000	348

Figure Example

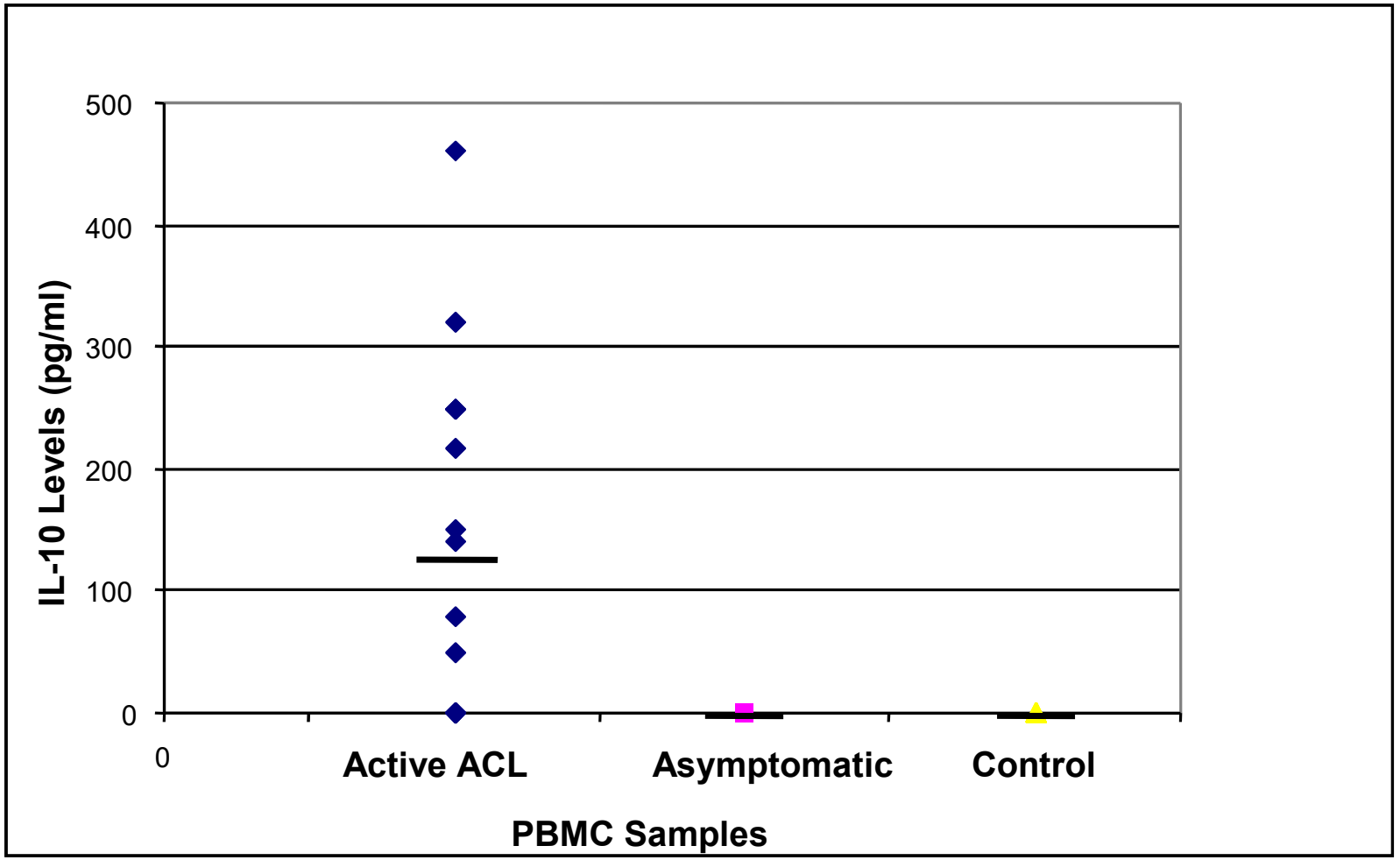


Figure 1. IL-10 produced by PBMCs in response to stimulation with the Leishmania antigen.
Peripheral blood mononuclear cells (PBMCs) collected from people with active atypical cutaneous leishmaniasis (ACL) infection, people with asymptomatic ACL, and uninfected people (control) were stimulated with 2 mg of soluble Leishmania antigen (SLA). IL-10 levels were measured by ELISA.

Graphing a small study

- Beware of presenting a small amount of data in graphs
  - Showing data in graphs can be a dramatic and effective way to show an effect or trend but if your data set is too small you can't say that you are seeing a potential trend or a real effect. So, showing it in a graph can be misleading.

- If you do use a graph with a small number of samples, clearly state how many data points you used

Figure Example

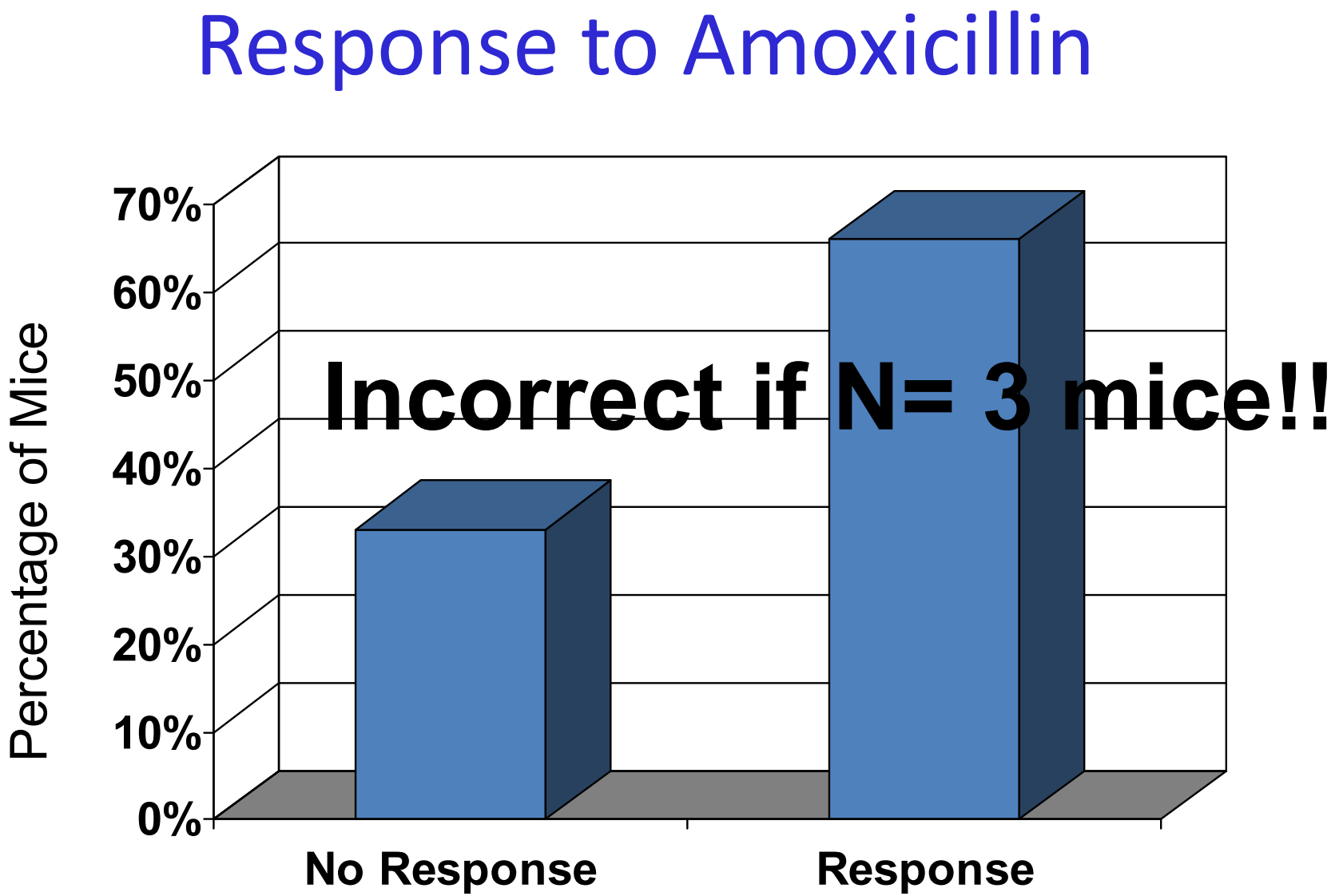
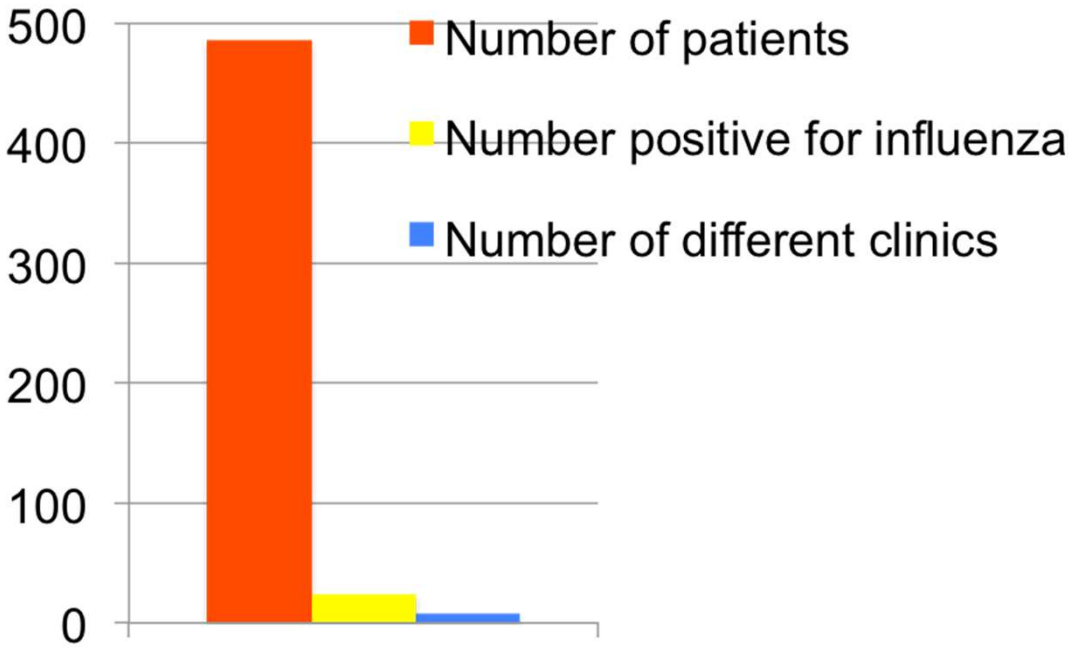


Figure 1. Percentage of mice that responded to amoxicillin treatment. Three mice were treated with 0.5 mg/ml amoxicillin for 7 days.

Unnecessary figure

Beware of presenting data that can be easily written in 1 sentence in text.



The figure presents the data of 486 patients tested for influenza at 8 different clinics where 24 of the 486 were positive for influenza. But the figure provides little in the way of details that could be included in 1 sentence in the text.

Each figure or table should stand alone

- Be understandable by itself
  - Should not have to read the text
  - Its title and descriptions should be enough

Ref: V. McMillan

Figure Options

- Graph → Data in connected series
- Chart → Data in separate series
- Picture → Must be seen (Photos)
- Diagram → Model to show concepts

Graphs

- Use to show a trend or pattern
- Generally a graph is not necessary when trends or relationships are not statistically significant
- If a cause and effect relationship:
  - X axis is the independent variable
  - Y axis is the dependent variable

Ref: V. McMillan

Recall this example graph

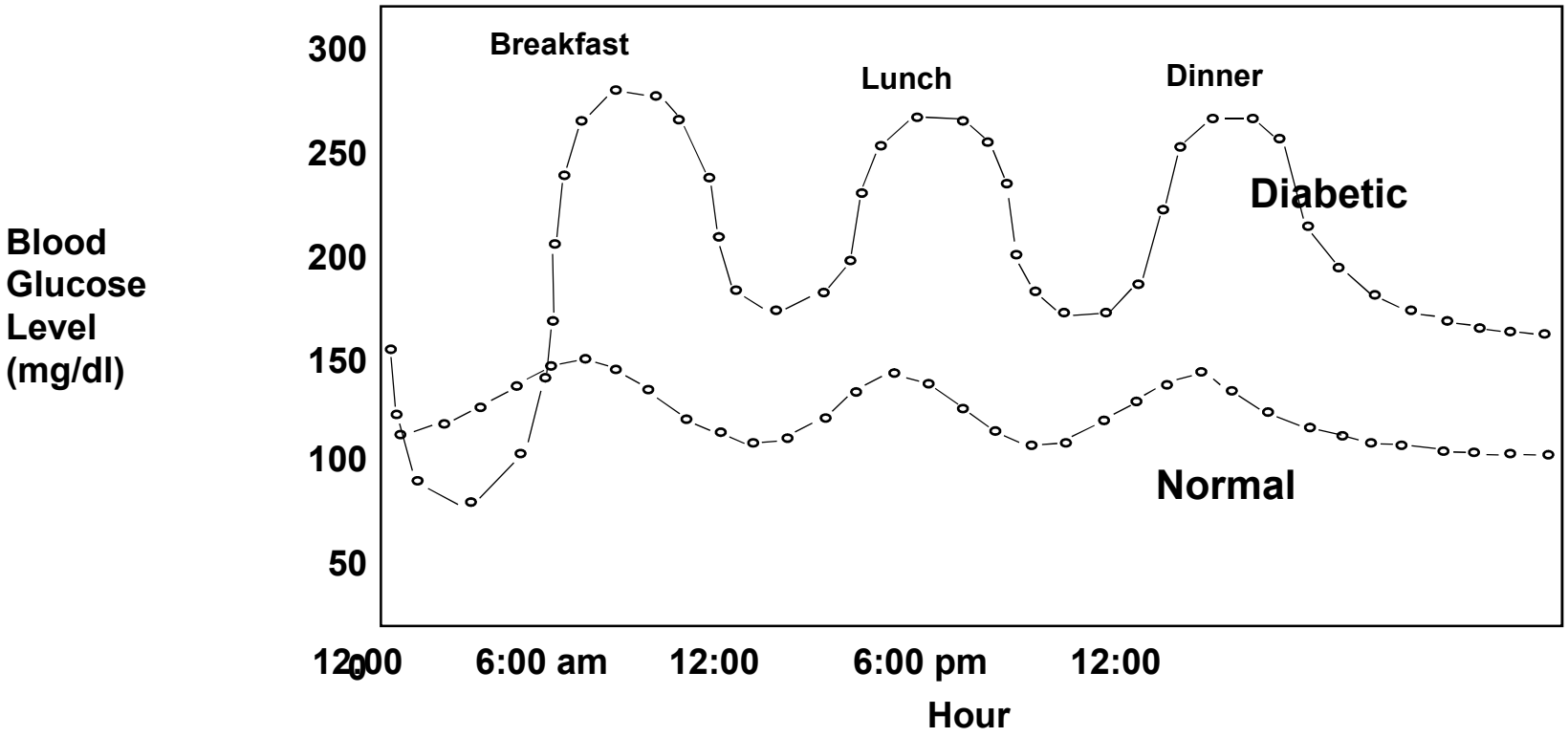


Figure 11. Blood glucose levels over time for normal individual and diabetic subjects

Another example graph

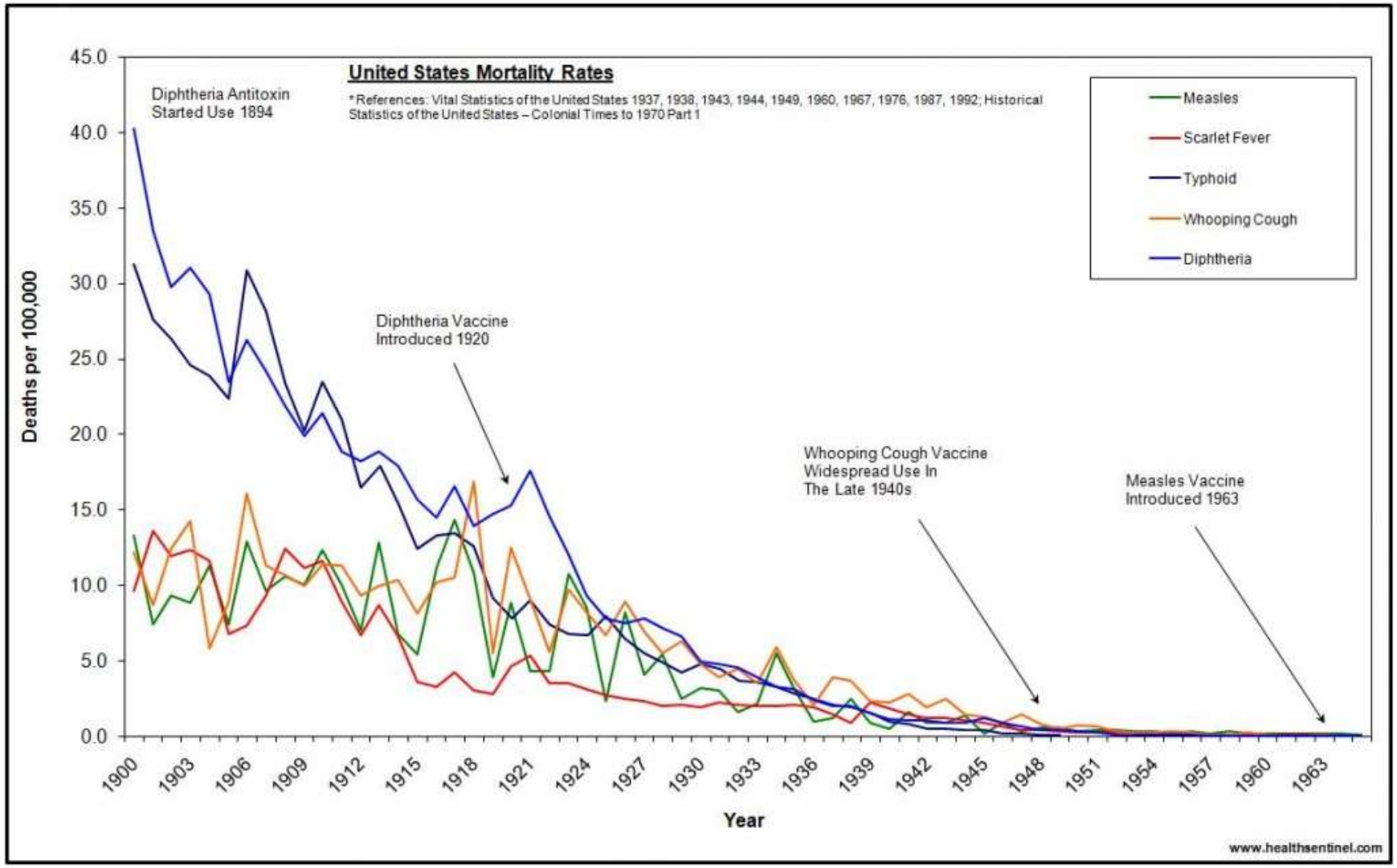


Figure Example

Another type of graph

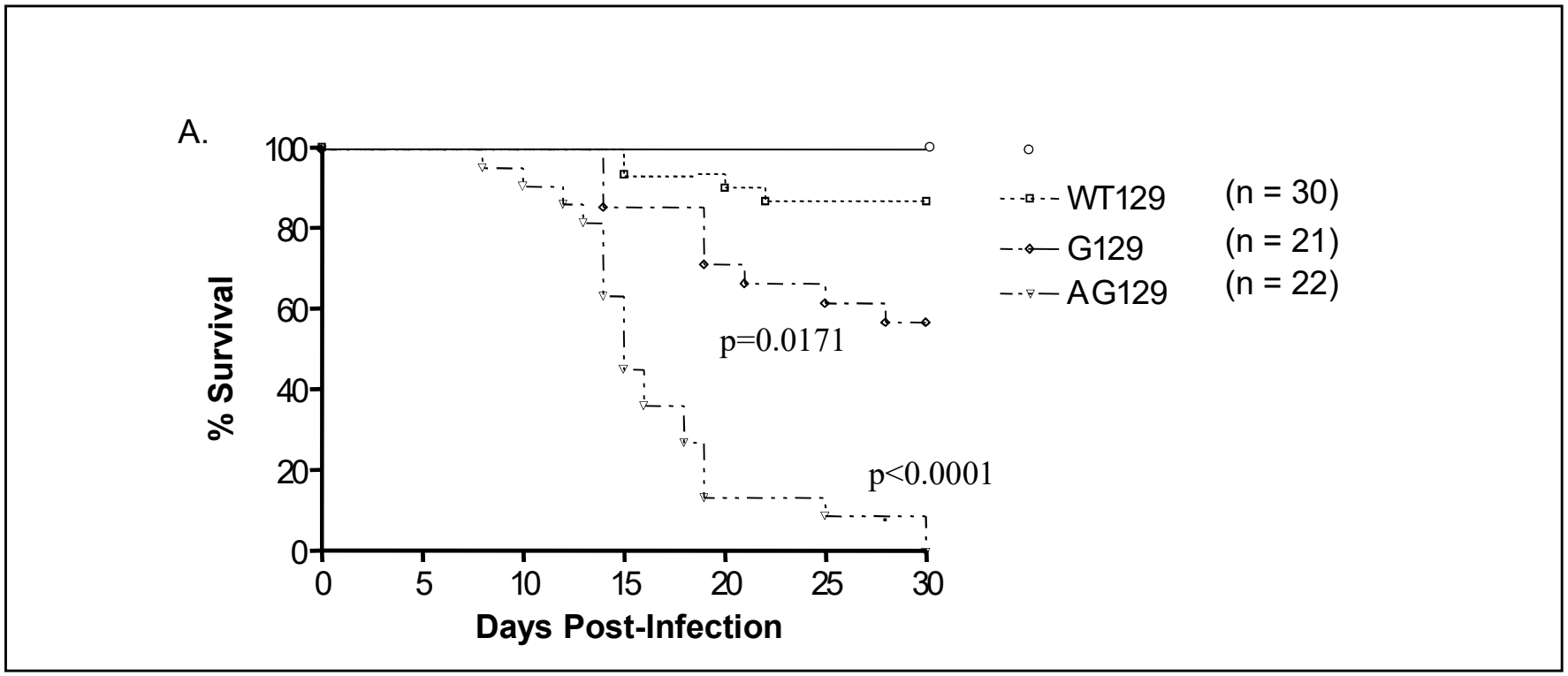


Figure 1. Survival of mice lacking the γ IFN and/or α/β IFN receptor genes following infection with dengue virus. A wild type mouse strain (WT129), a transgenic strain deficient for the gamma IFN receptor only (G129), as well as a transgenic strain deficient for the alpha/beta and gamma IFN receptor genes in combination (AG129) were infected with 10^8 PFU of dengue virus (DEN2). Survival over time was determined. The statistical significance (p-values) of the differences in survival between the transgenic and wildtype strains are provided.

Figure Example

Charts: Graphical but not connected

Figure 1. Mechanisms causing NHL concussion.
The causes of NHL concussion or suspected concussion were documented for the subset of injuries occurring during the 10 randomly selected weeks for all 3 seasons. The proportion of injuries within each season caused by each mechanism is shown, with the number of injuries above each bar. Unintentional actions included tripping and colliding with a teammate. The rates of each mechanism remained constant over the seasons tested ($p>0.05$ for all).

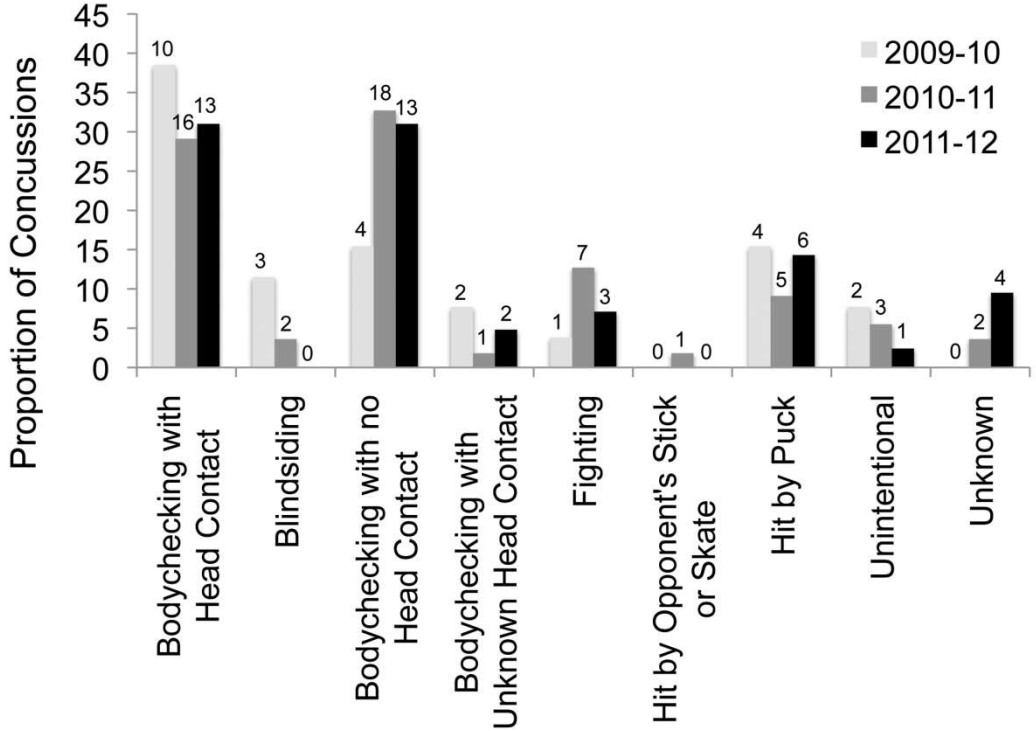
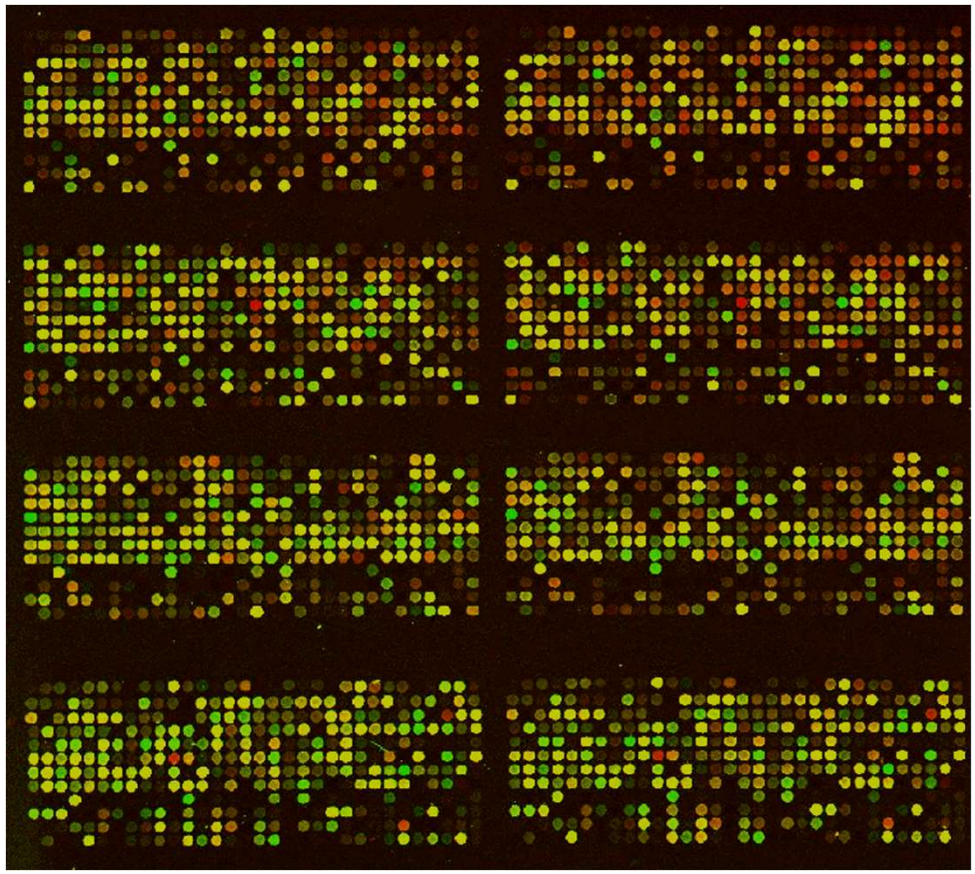


Figure Example

From raw data to figure (chart)



Microarray data looking at RNA expression for thousands of genes on chip.
Red =up regulation
Green= down regulation

How do you present this data?

Figure Example

Pie-charts of microarray data

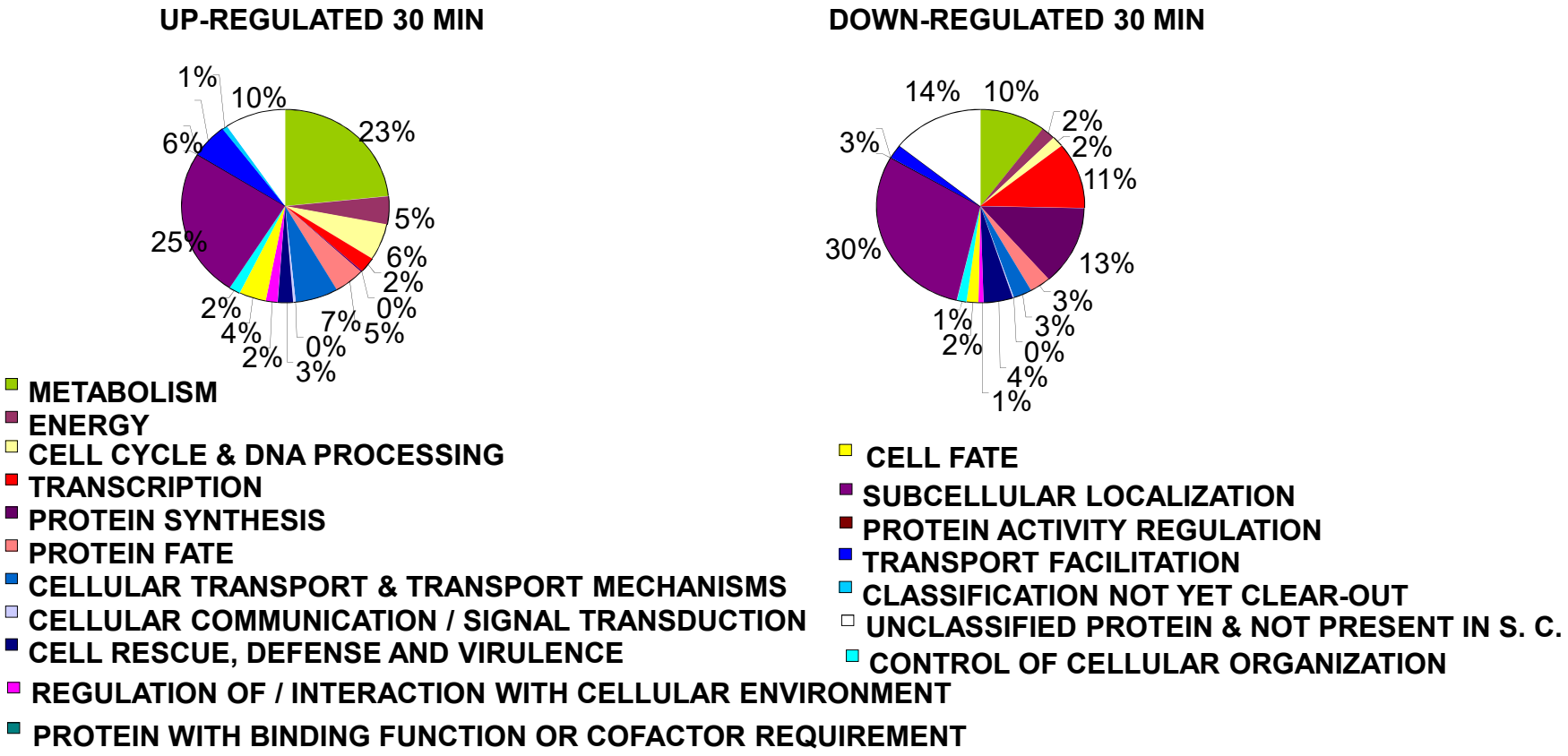


Figure 1. Effect of cycloheximide upon gene expression in *Saccharomyces cerevisiae*. The wildtype yeast strain (YAS1180) was grown in the presence of 50 ng/ml cycloheximide for 30 minutes at 30°C. Cells were harvested and total RNA was extracted. RNA expression of different classes of genes was determined using the Affymetrix GeneChip Expression System for Saccharomyces cerevisiae.

Figure Example

More microarray data

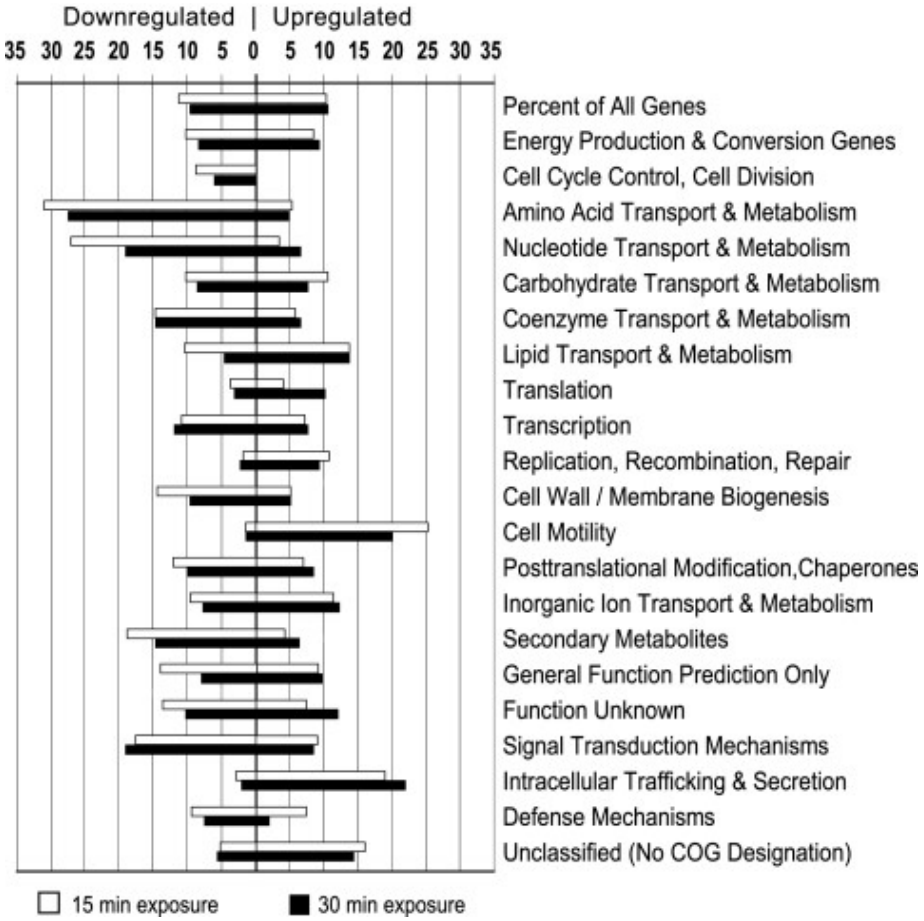
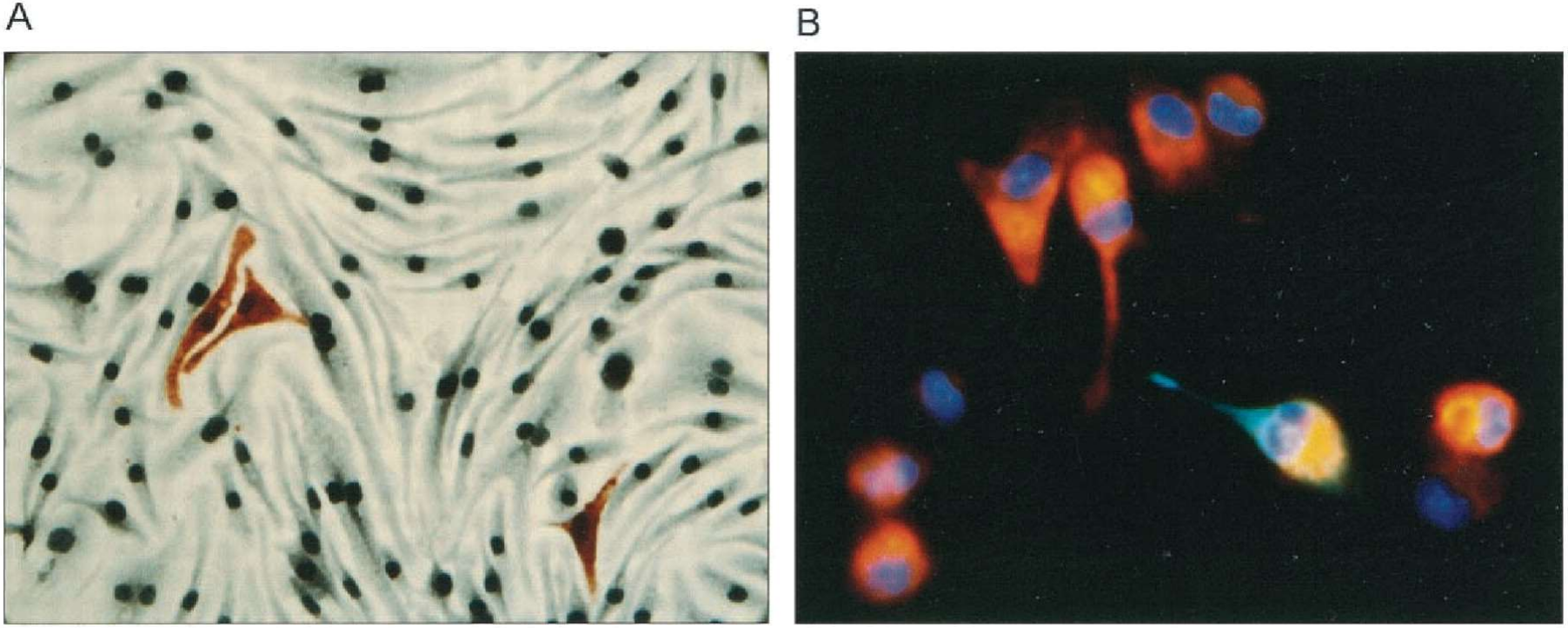


Figure X. Grouping into functional categories of differentially regulated genes in ECO157 in romaine lettuce lysates compared to M9-glucose medium, as determined by microarray analysis. The categories of orthologous genes (COG) were used for grouping. Bars represent the percentage of genes with decreased or increased expression in a given category after a 15-min or a 30-min exposure to lettuce lysates.

Kyle et al 2008

Pictures

When nothing else will do



Dengue virus (DENV)—2 infection in bone marrow—derived macrophages (BMDMs) from AG129 mice (A) and peritoneal macrophages from A129 mice (B).

Kyle et al 2007

Pictures

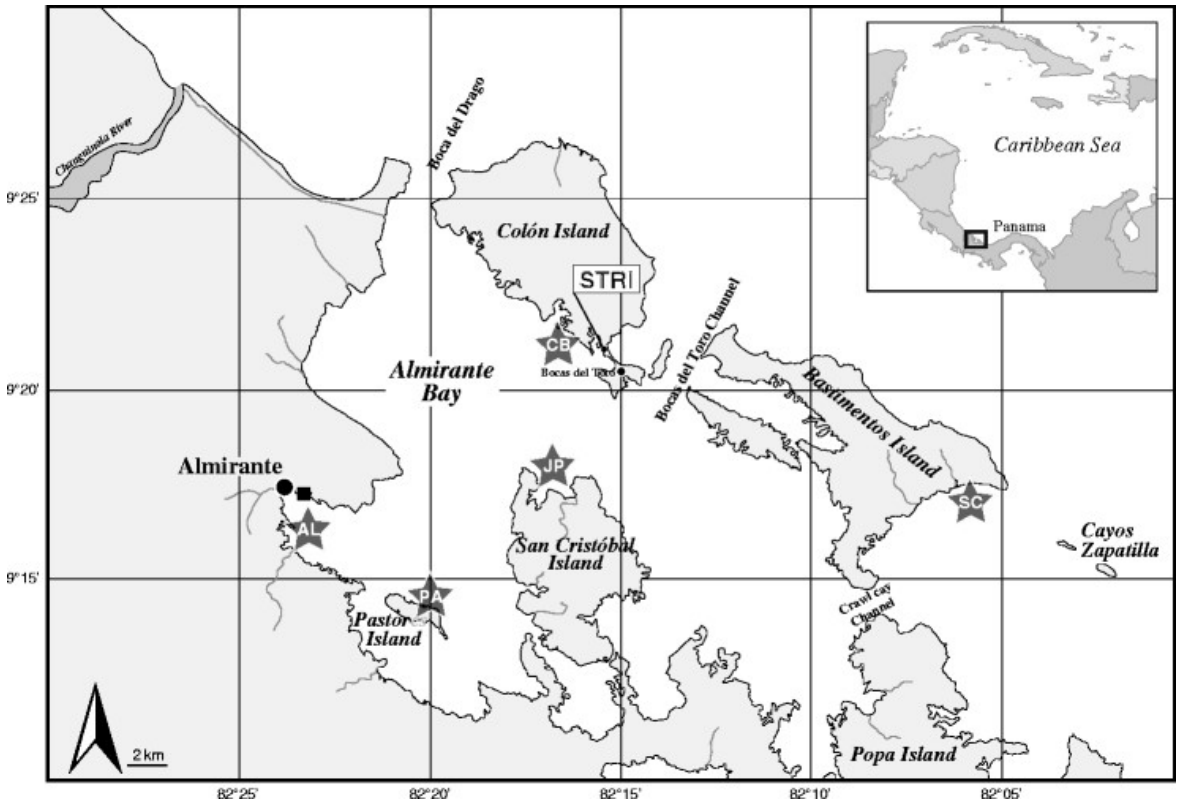


Fig. 1 Survey sites within the Bocas del Toro Archipelago are depicted by grey stars. AL = Almirante, PA = Pastores, JP = Juan Point, CB = Casa Blanca, SC = Salt Creek, STRI = Smithsonian Tropical Research Institute. Commercial shipping vessels enter the bay through the Boca del Drago inlet in the north and proceed to the Port of Almirante (black square). The town of Almirante is represented by the black circle .

Figure Example

Gels

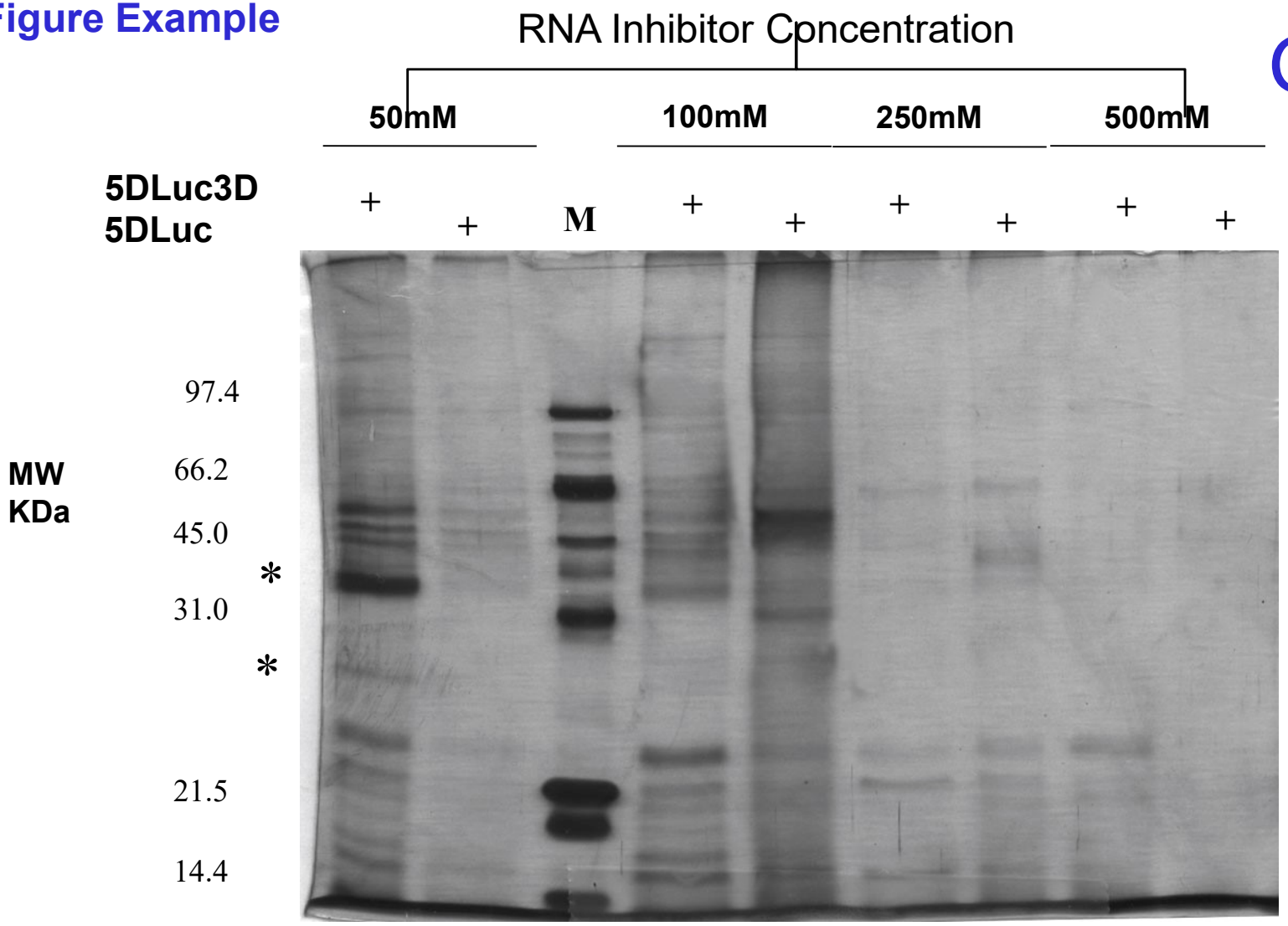


Figure 1. Effect of dengue virus 5' and 3' UTRs on protein translation in the presence and absence of inhibitor. Luciferase reporter constructs containing both the dengue virus 5' and 3' UTR (5DLuc3D) or the dengue virus 5'UTR only (5DLuc) were transfected into cells. Cells were subsequently exposed to various concentrations of an RNA inhibitor. Cell extracts were harvested and analyzed by SDS-Gel electrophoresis followed by Coomassie blue staining. Luciferase protein products of 28 Kda and 43 Kda are noted with an *.

Diagrams

To show complex relationships

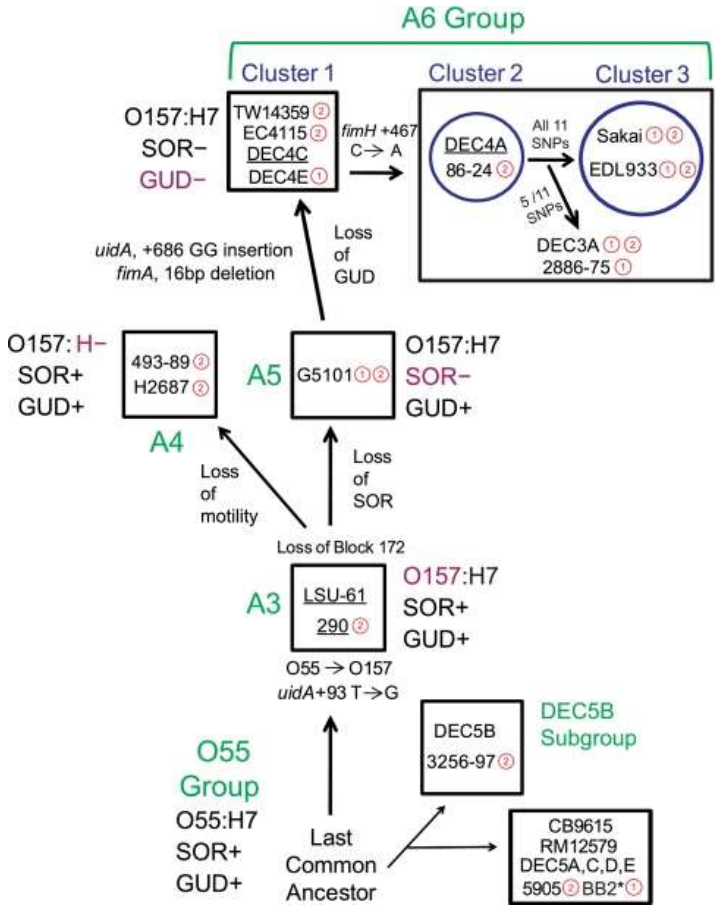
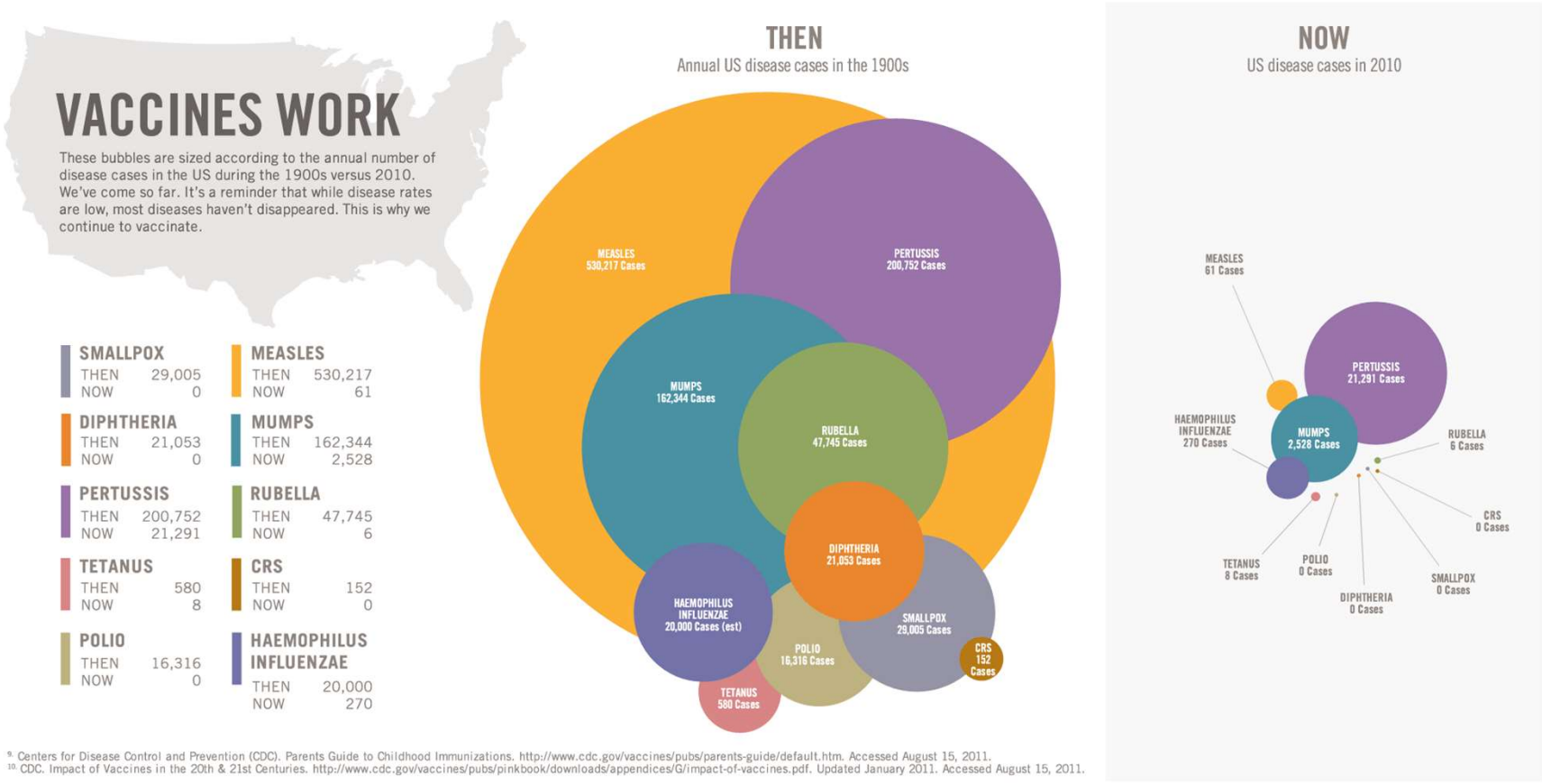


Figure X. Revised evolutionary model describing evolution of O55:H7 to O157:H7 with variable *stx* presence. A comprehensive evolutionary model is presented, encompassing elements of previous models as well as two new O55:H7 clades based on genetic differences identified in this study. Representative strains for each group have been placed within each square or circle, and the presence of *stx* phages is noted for each representative strain in red circles. Nonhuman isolates have been underlined. 1, stx_1 positive; 2, stx_2 positive; *, single known sorbitol-negative O55:H7 strain.

Kyle et al 2012

My favorite graphic



[9] Centers for Disease Control and Prevention (CDC). Parents Guide to Childhood Immunizations. http://www.cdc.gov/vaccines/pubs/parents-guide/default.htm. Accessed August 15, 2011.
[10] CDC. Impact of Vaccines in the 20th & 21st Centuries. http://www.cdc.gov/vaccines/pubs/pinkbook/downloads/appendices/G/impact-of-vaccines.pdf. Updated January 2011. Accessed August 15, 2011.

Summary of Figure formats

- Different types
  - Graph
  - Chart
  - Picture (gels, flow cytometry)
  - Diagram
- All have
  - Figure Legend (title and a brief description of experiment)

Four parts to figure legends

1. Title
   - One sentence to identify the main point of the figure.
2. Brief experimental details
   - Enough details so that the reader can understand figure.
3. Definitions
   - Symbols or bar patterns that that are not explained in figure.

Antigen present

Control

4. Statistical information
   - Number of samples, p-values, etc.

Ref: V. McMillan

Tables

- Use a table to present many numerical values
- Don’t pack in too much information!
- Don’t include columns that have the same value throughout. You can include that information in a caption or in the text.

Ref: V. McMillan

Table format

- Columns and rows
  - Organize a table so that the similar items read down, not across
- Table title
- Footnotes
- Look at the format in other papers as a guideline

Ref: V. McMillan

Table Example

TABLE 1. Baseline differences in sociodemographic characteristics by treatment group, Ecuador, Panama, and Uruguay, 2002–2005

	Amalgam		Autramatic restorative treatment	
Characteristic	No.	%	No.	%
Gender				
Male	368	49.1	418	48.1
Female	393	50.8	450	51.8
Age (years)				
7	89	12.2	139	17.3
8	312	38.2	350	37.2
9	360	48.5	389	45.4
Country				
Ecuador	344	56.5	391	56.4
Panama	269	18.6	331	36.8
Uruguay	148	24.8	174	24.4
Geographic location				
Urban	465	62.2	537	63.1
Rural	296	37.7	331	36.8

Table Example

Table 2. Concussion incidence in the OHL and NHL by season.

		Concussions			+ Suspected		
	Season	**Incidence/100 games**	**IRR**	**95% CI**	**Incidence/100 games**	**IRR**	**95% CI**
NHL	2009–10	3.58	0.64*	0.42, 0.96	6.26	0.61*	0.45, 0.83
NHL	2010–11	5.28	1	N/A	9.76	1	N/A
NHL	2011–12	6.83	1.35	0.96, 1.89	10.24	1.05	0.80, 1.38
OHL	2009–10	3.38	0.60	0.35, 1.05	4.71	0.65	0.40, 1.07
OHL	2010–11	5.59	1	N/A	6.91	1	N/A
OHL	2011–12	6.18	1.18	0.74, 1.89	9.71	1.54*	1.02, 2.31

Incidence rates were calculated per 100 regular season games. IRRs for concussions and concussions plus suspected concussions were calculated relative to the 2010–11 season. Concussion incidence rate in the NHL was lower in 2009–10 than in 2010–11 ($p = 0.002$ for concussion and suspected concussion, $p = 0.029$ for concussion), but there was no significant difference between 2010–11 and 2011–12 ($p = 0.727$ for concussion and suspected concussion, $p = 0.086$ for concussion). OHL concussion incidence rates were not different between 2009–10 and 2010–11 ($p = 0.074$ for concussion and suspected concussion, $p = 0.09$ for concussion) but concussions and suspected concussion increased from 2010–11 to 2011–12 ($p = 0.039$, $p = 0.483$ for concussion only).

*indicates $p<0.05$ within each league relative to 2010–11.

doi:10.1371/journal.pone.0069122.t002

Table footnotes

- Footnotes are BRIEF explanations about data including
  - Exceptions
  - Abbreviations
  - Statistics
    - p-values for data were <0.05.

- Do not write out information that belongs in the results!

Table Example

TABLE 3. Odds ratios of failures of alternatives to amalgam treatment at 12 months, Panama, Ecuador, and Uruguay, 2002–2005

Atraumatic restorative treatment	Group only	Group and country	Group, country, and age	Group, country, age, and sex[a]	Group, country, age, and sex weighted clustering
By dentist	1.88[b]	1.81[b]	1.80[b]	1.81[b]	1.75[b]
By auxiliary	4.19[b]	4.10[b]	4.07[b]	4.18[b]	3.43[b]

[a] Results are insensitive to clustering by child with multiple teeth or by operator who treated multiple teeth.
[b] $P < 0.05$.