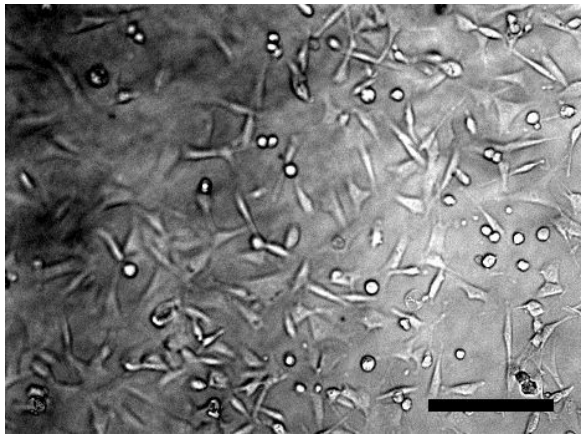
Basic Statistics; Standards in Scientific Communities I

Module 3, Lecture 3

20.109 Spring 2011

Lecture 2 review

- What properties of hydrogels are advantageous for soft TE?
- What is meant by bioactivity and how can it be introduced?
- What are the two major matrix components of cartilage and how do they support tissue function?

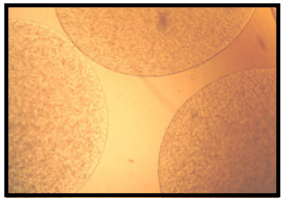


Topics for Lecture 3

- Module 3 so far, and Day 3 plan
- Introduction to statistics
  - confidence intervals
  - t-test
- Standards in scientific communities
  - general engineering principles
  - standards in synthetic biology
  - standards in data sharing

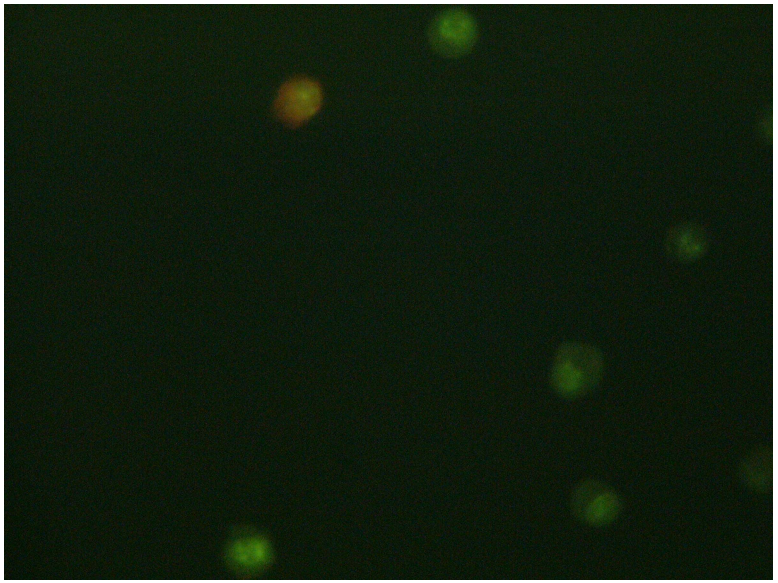
Module progress: week 1

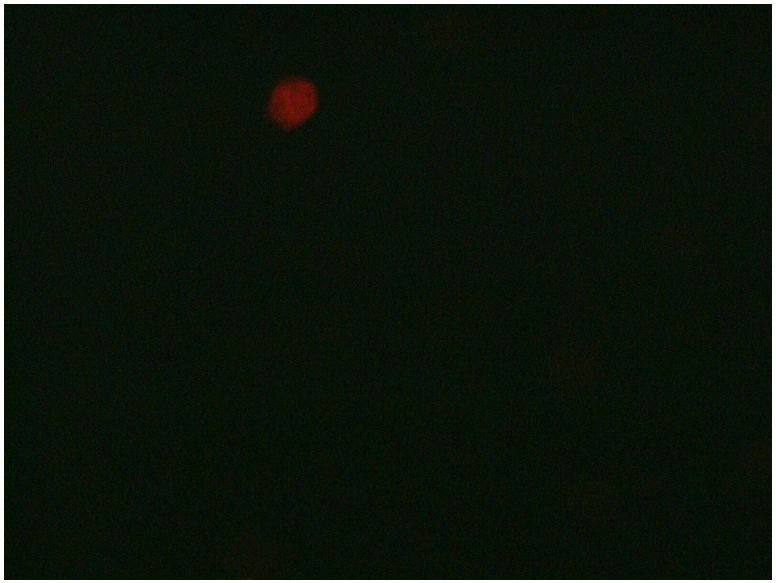
- Day 1: culture design
  - What did you test?



- Day 2: culture initiation
  - Cells receiving fresh media every 2 days

Module day 3: test cell viability





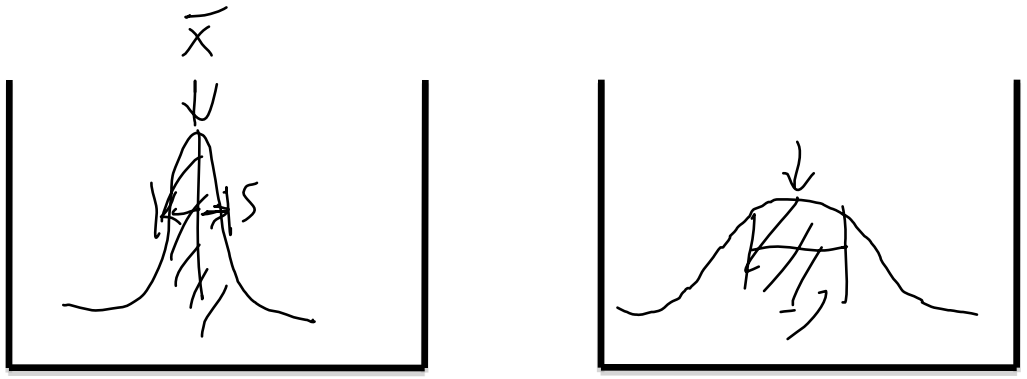
Green stain: SYTO10 = viability
Red stain: ethidium = cytotoxicity

} Assay readout: fluorescence

Working principle? **Relative cell-permeability**

Statistics review: basics

- Essential concepts: standard deviation (*s*), mean ($\bar{x}$), sample size *n,* degrees of freedom *DOF*
- Normal (Gaussian) distribution



1 *s* includes

68 %

of the data

x-axis: measured values (intensity)

y-axis: # of samples w/ that value

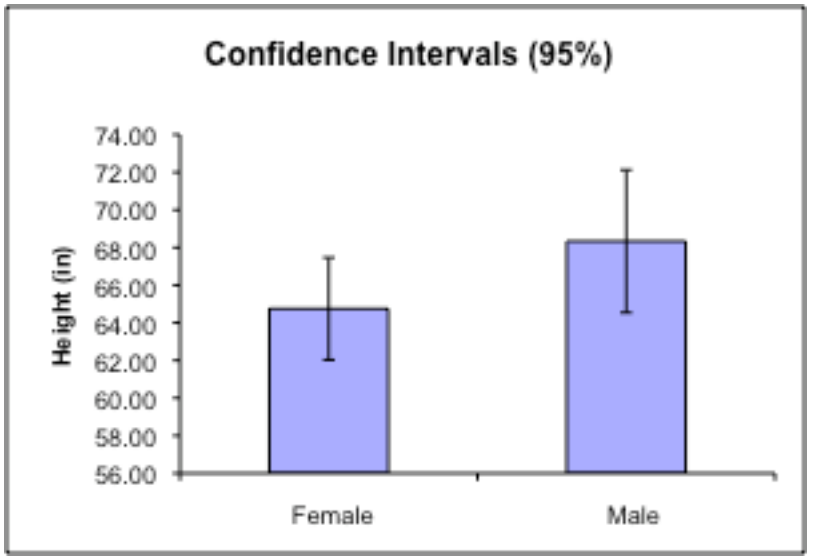
Confidence intervals (CI): principle

- $\overline{x}$ = 60 (sample/measured mean)
- 95% CI calculated to be ± 3 from real data
- Thus: 95% of the time our population (true) mean μ lies in the range, here 60 ± 3
  - subtly different from 95% likely that the range 60 ± 3 contains the population (true) mean μ, which we can't say
- 90% CI: $\mu = \overline{x} \pm a$ where $\boldsymbol{a < 3}$ $a > 3$ $a = 3$?

  trade-off between precision and confidence
- Consider betting example
- What about n? **as n increases, more precise**

Calculating confidence intervals (CI)

$$\mu = \bar{x} \pm \frac{t\,s}{\sqrt{n}}$$



- *t* is tabulated by DOF vs CI%
  - DOF = *n* - 1
- In Excel, us *TINV* function
  - input *p*-value = (100-CI)/100

why? $\sum(\text{errors} = x_n - \bar{x}) = 0$ → constraint

0.L = 95%, $p = 0.05$

Introduction to t-test

- Every statistical test
  - has assumptions
  - asks a specific question
  - requires human interpretation
- Some t-test assumptions
  - normal distribution (cf. Mann-Whitney test)
  - equal variances (type 2 in Excel; type 3 unequal)
- Posing a question *are average male and female heights different at a confidence level of 95%*

Calculating t-test significance

$$t_{calc} = \frac{\bar{x}_1 - \bar{x}_2}{s_{pooled}} \sqrt{\frac{n_1 n_2}{n_1 + n_2}}$$

DOF = $n_1 + n_2 - 2$

t_{table} listed by DOF vs. C.L.

- If t_{calc} > t_{table} difference *is* significant at that C.L.
- In Excel, us *TTEST* function
- Excel returns *p*-value → confidence level (CL)
- 1-tailed vs. 2-tailed test

$p = 0.01$, C.L. 99%

1- one-sided hypothesis in advance

2- no a priori hypothesis

Assignment for report

- Get live cell count and/or live cell percent values for both culture conditions
- Calculate 95% CI for both means
- Plot means on bar graph with CI error bars
- Apply t-test to the means
  - For multiple comparisons, ANOVA is better
  - Comparing many means requires correction
  - Remember, $p = 0.05$ means 1 in 20 false positives!

Interlude: what (if any) should be off-limits to science?

1. "Hallucinogens have doctors tuning in again" *NY Times* April 2010

Researchers from around the world are gathering this week in San Jose, Calif., for the largest conference on psychedelic science held in the United States in four decades. They plan to discuss studies of psilocybin and other psychedelics for treating depression in cancer patients, obsessive-compulsive disorder, end-of-life anxiety, post-traumatic stress disorder and addiction to drugs or alcohol.

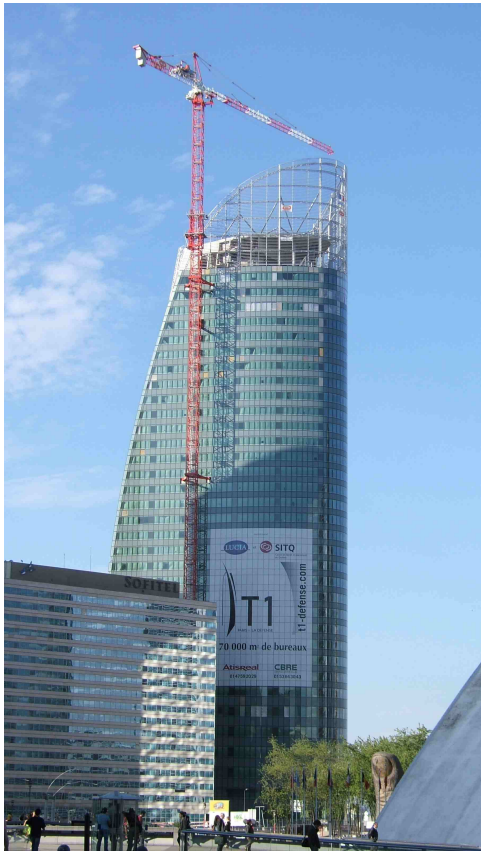
2. "Towards responsible use of cognitive-enhancing drugs by the healthy" *Nature* ***456, 702-705***

Thinking critically about module goals

- Purpose of experiment
  - Local compare 2 culture conditions ⇒ effect on cell phenotype
  - Global cartilage regeneration
- All well and good, but…
- Can we move beyond empiricism – tissue *engineering*
- E.g., broadly useful biomaterials
  - goal: control degradability over wide range
  - "a lot of chemical calculations later, we estimated that the anhydride bond would be the right one"
  - Robert Langer, *MRS Bulletin* **31**(2006).

Engineering principles, after D. Endy

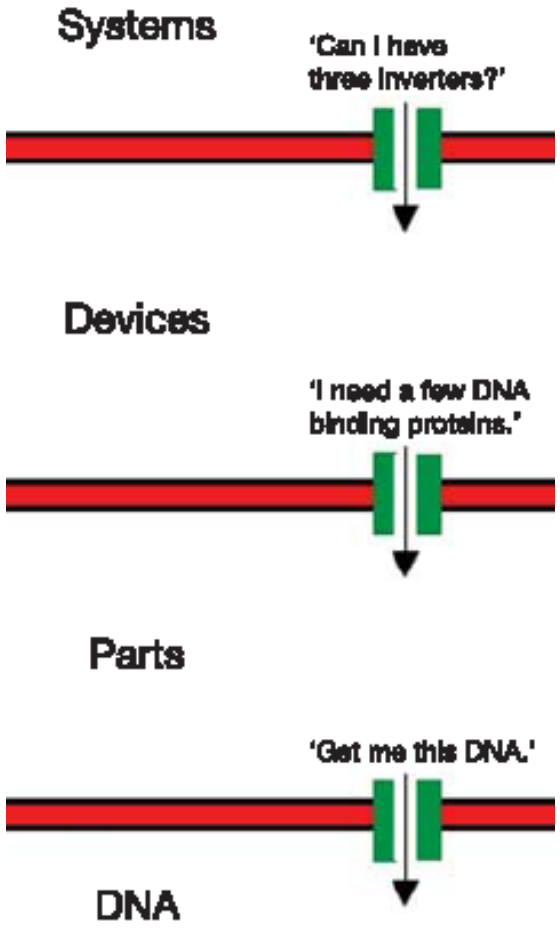
- D. Endy, *Nature* **438**:449 (2005)
- Is biology too complex to engineer, or does it simply require key “foundational technologies”?
- Systematic vs. *ad hoc* approach
- Abstraction
  - e.g., software function libraries
- Decoupling
  - e.g., architecture vs. construction
- Standardization
  - screw threads, train tracks, internet protocols
- Given your Module 2 experience, what should we standardize to engineer biology?



Public domain image (Wikimedia Commons)

Application to synthetic biology

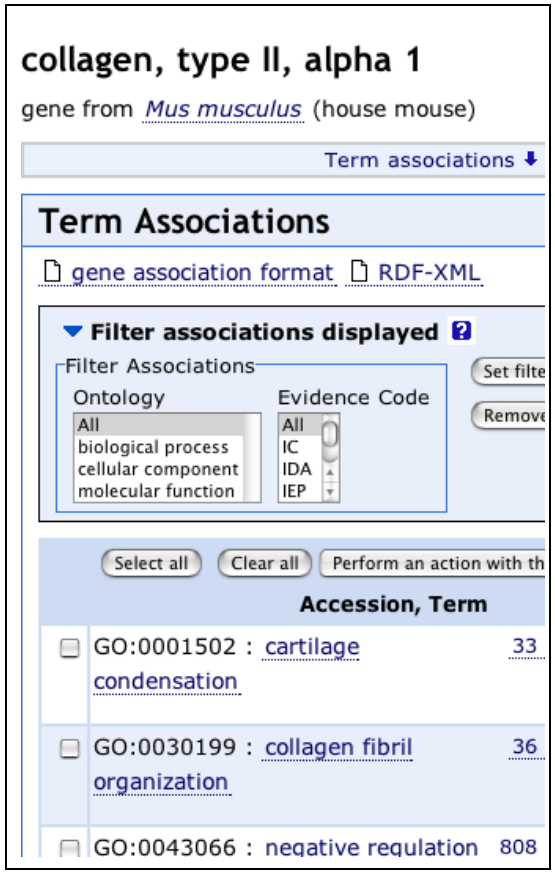
- D. Endy, *Nature* **438**:449 (2005)
- Recall, synthetic biology = "programming" cells/DNA to perform desired tasks
- Abstraction
  - DNA → parts → devices → systems
  - materials processing to avoid unruly structures
- Decoupling
  - DNA design vs. fabrication (rapid, large-scale)
- Standardization
  - functional (e.g., RBS strength)
  - assays
  - system conditions
  - standard junctions to combine parts



From D. Endy, *Nature* **438**:449

Data standards: what and why?

- Brooksbank & Quackenbush, *OMICS*, **10**:94 (2006)
- High-throughput methods are data-rich
- Standards for **collection** and/or **sharing**
- Reasons
  - shared language (human and computer)
  - compare experiments across labs
  - avoid reinventing the wheel
  - integration of information across levels
- Examples
  - MIAME for microarrays
  - Gene Ontology (protein functions)
- Who drives standards?
  - scientists, funding agencies, journals, industry

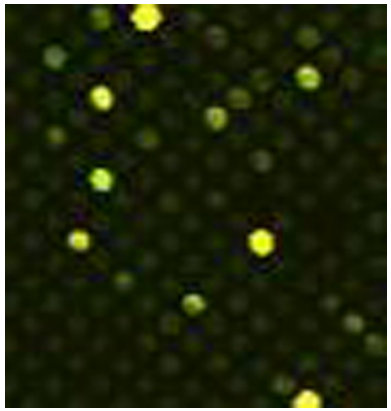


www.geneontology.org

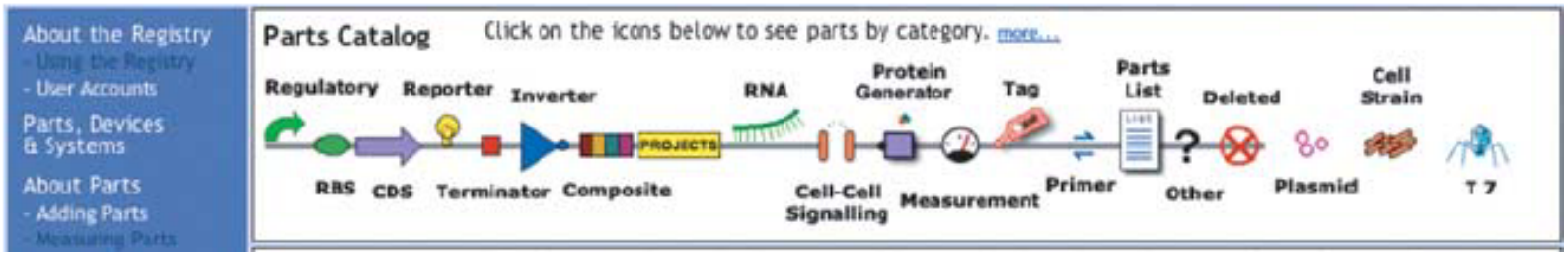
Lecture 3: conclusions

- Confidence intervals and t-tests are two useful statistical concepts.
- Standardizing data sharing and collection is of interest in several BE disciplines.

Microarray data



From D. Endy, *Nature* **438**:449 (standardized biological "parts")



Next time: *discussion* of standards in TE; more about cell viability and microscopy