Design of Experiment

Yi-Ju Li, Ph.D.

High-throughput Sequencing Workshop

Department of Biostatistics & Bioinformatics
Duke University Medical Center

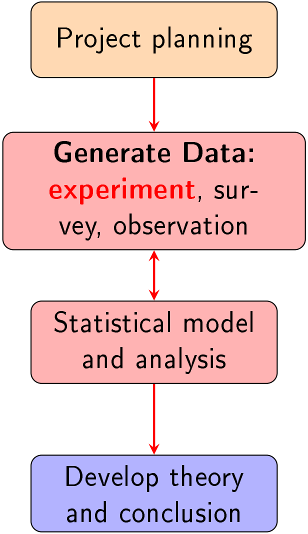
July 5, 2018

Outline

- Definition and Principles of Design of Experiment (DOE)
- Basic statistics
- Types of experimental designs for basic science research
- Power calculation for sample size
- DOE consideration for RNA-Seq

Definition and Principles

General study workflow



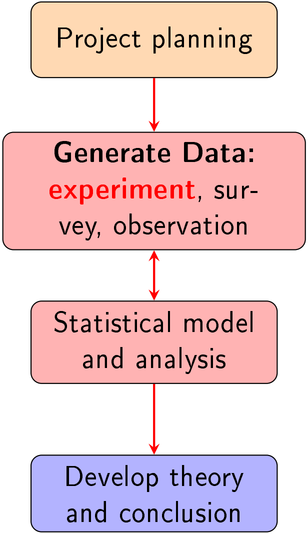
Project planning

Generate Data: **experiment**, survey, observation

Statistical model and analysis

Develop theory and conclusion

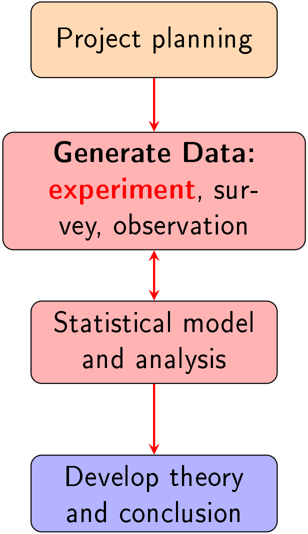
General study workflow


Project planning
Generate Data: **experiment**, survey, observation
Statistical model and analysis
Develop theory and conclusion

Project planning

Hypothesis; what to be measured; influential factors

General study workflow


Project planning
Generate Data: experiment, survey, observation
Statistical model and analysis
Develop theory and conclusion

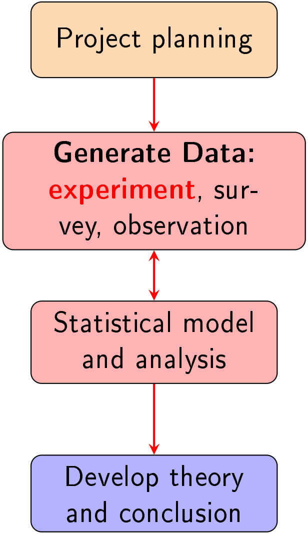
Project planning

Hypothesis; what to be measured; influential factors

Experimental studies

Ability to control the source of variability

General study workflow


Project planning
Generate Data: **experiment**, survey, observation
Statistical model and analysis
Develop theory and conclusion

Project planning

Hypothesis; what to be measured; influential factors

Experimental studies

Ability to control the source of variability

Observational studies

No controls over the source of variability

Basic definition of design of experiment (DOE)

- **Experiment:** A process that generates data to achieve specific objective
- **All data are subject to variation.**
- **DOE:** A systematic method to determine the effect of a factor(s) to the outputs (responses) of the experiment based on predefined questions (*e.g.* , hypothesis, theory, model). An effective experiment can
  - eliminate known sources of bias
  - prevent unknown source of bias
  - obtain data with high accuracy and precision.
- R.A. Fisher pioneered the field of statistical principals of experimental design.

Main elements in EOD

- **Formulate research questions and hypothesis.**
- **Experimental units:** The entities that experimental procedures are applied to.
  - Examples: Mice, plants, patients, etc.
  - Need to be representative for the inference to be made.
- **Observation units or response variables:** Any outcomes or results of the experiment (*e.g.* . gene expression of the RNA-Seq study)
  - Responses are only comparable if they are measured from homogeneous experimental units.

More on main elements

- **Factors:** Variables to be investigated to determine its effect to the response variable (*e.g.* treatment effect)
  - It should be defined prior to the experiment.
  - It can be controlled by experimenter.
- **Effect**: Changes in the average response between levels of a factor, or between two experimental conditions.
- **Covariate:** May affect the response but cannot be controlled in an experiment.

More on formulating hypothesis

1. Establish a study objective from a given scientific question.
2. Translate study objective to a testable hypothesis
3. **Null hypothesis:** No measurement differences or factor effects between groups
4. **Alternative hypothesis:** Certain measurement differences or factor effects between groups
    - Mostly it is the goal you want to achieve in your study objective.

Examples: From study objective to hypothesis

1. **Study objective:** ’To examine the complications, mortality, cost and discharge status of patients with disease X’

Examples: From study objective to hypothesis

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

   Concerns:

   - Examine = estimate rates? or Examine = compare rates?
   - There are no comparable groups, so we can't establish a testable hypothesis.

Examples: From study objective to hypothesis

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

   Concerns:

   - Examine = estimate rates? or Examine = compare rates?
   - There are no comparable groups, so we can't establish a testable hypothesis.

2. **Study objective:** 'To identify differential expression genes between E *Coli* stressed by high and neutral pH level'

Examples: From study objective to hypothesis

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

   Concerns:

   - Examine = estimate rates? or Examine = compare rates?
   - There are no comparable groups, so we can't establish a testable hypothesis.

2. **Study objective:** 'To identify differential expression genes between E *Coli* stressed by high and neutral pH level'
   Hypothesis:

Examples: From study objective to hypothesis

1. **Study objective:** 'To examine the complications, mortality, cost and discharge status of patients with disease X'

   Concerns:

   - Examine = estimate rates? or Examine = compare rates?
   - There are no comparable groups, so we can't establish a testable hypothesis.

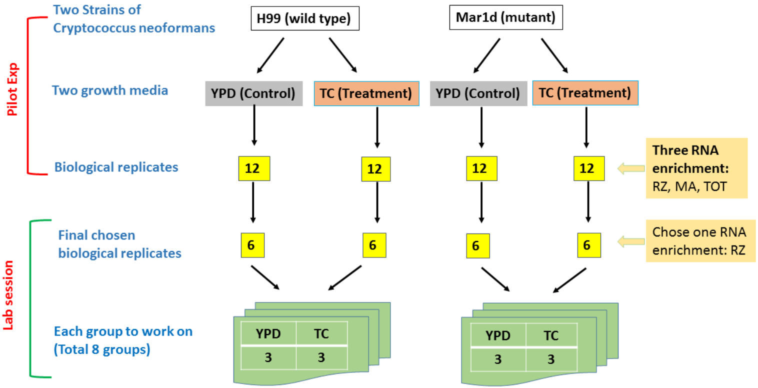
2. **Study objective:** 'To identify differential expression genes between E *Coli* stressed by high and neutral pH level'

   Hypothesis:

   - **Null hypothesis:** There is no difference in expression level of 'a gene' between pH conditions; $\mu_1 = \mu_2$, μ for average gene expression level.
   - **Alternative hypothesis:** There is difference in expression level of 'a gene' between pH conditions; $\mu_1 \neq \mu_2$.

Experiment for this workshop

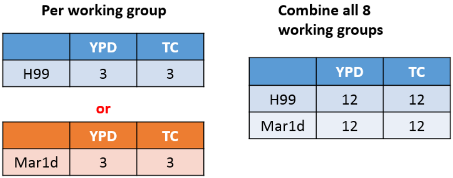
A two-factor experiment for Cryptococcus neoformans (fungus):


Pilot Exp
Two Strains of Cryptococcus neoformans
Two growth media
Biological replicates
Lab session
Final chosen biological replicates
Each group to work on (Total 8 groups)
H99 (wild type)
Mar1d (mutant)
YPD (Control)
TC (Treatment)
YPD (Control)
TC (Treatment)
12
12
12
12
Three RNA enrichment: RZ, MA, TOT
6
6
6
6
Chose one RNA enrichment: RZ
YPD
TC
3
3
YPD
TC
3
3

RZ: ribozero rRNA depletion; MA: polyA enrichment; TOT: total RNA

Practice

Experiment: RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



Per working group

	YPD	TC
H99	3	3

or

	YPD	TC
Mar1d	3	3

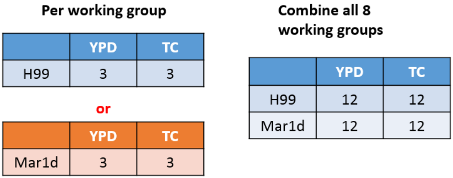
Combine all 8 working groups

	YPD	TC
H99	12	12
Mar1d	12	12

- Study objective?

Practice

Experiment: RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



Per working group

	YPD	TC
H99	3	3

or

	YPD	TC
Mar1d	3	3

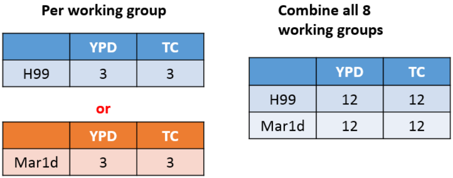
Combine all 8 working groups

	YPD	TC
H99	12	12
Mar1d	12	12

- Study objective?
- Null and alternative hypotheses?

Practice

Experiment: RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



Per working group

	YPD	TC
H99	3	3

or

	YPD	TC
Mar1d	3	3

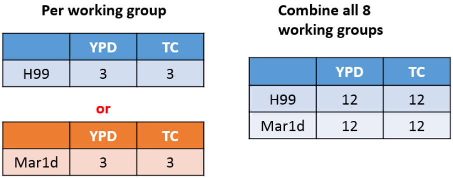
Combine all 8 working groups

	YPD	TC
H99	12	12
Mar1d	12	12

- Study objective?
- Null and alternative hypotheses?
- Experimental units?

Practice

Experiment: RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



Per working group

	YPD	TC
H99	3	3

or

	YPD	TC
Mar1d	3	3

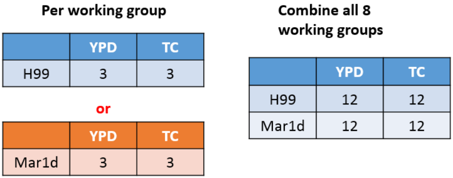
Combine all 8 working groups

	YPD	TC
H99	12	12
Mar1d	12	12

- Study objective?
- Null and alternative hypotheses?
- Experimental units?
- Observation units?

Practice

Experiment: RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



Per working group

	YPD	TC
H99	3	3

or

	YPD	TC
Mar1d	3	3

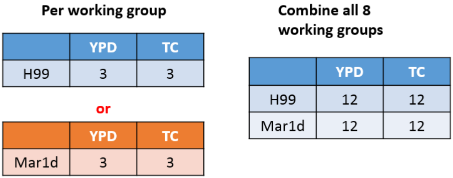
Combine all 8 working groups

	YPD	TC
H99	12	12
Mar1d	12	12

- Study objective?
- Null and alternative hypotheses?
- Experimental units?
- Observation units?
- Factors?

Practice

Experiment: RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



Per working group

	YPD	TC
H99	3	3

or

	YPD	TC
Mar1d	3	3

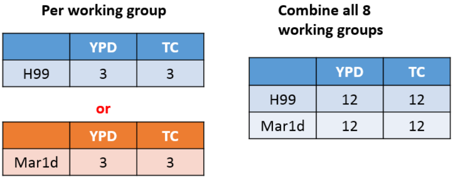
Combine all 8 working groups

	YPD	TC
H99	12	12
Mar1d	12	12

- Study objective?
- Null and alternative hypotheses?
- Experimental units?
- Observation units?
- Factors?
- Covariates?

Practice

Experiment: RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



Per working group

	YPD	TC
H99	3	3

or

	YPD	TC
Mar1d	3	3

Combine all 8 working groups

	YPD	TC
H99	12	12
Mar1d	12	12

- Study objective?
- Null and alternative hypotheses?
- Experimental units?
- Observation units?
- Factors?
- Covariates?

Common problems in experimental design

- Experimental variation may mask the factor effects.
  - For data with larger variation, it is more difficult to detect mean differences between two levels of a factor.
  - Sample size matters.
- Uncontrolled factors may compromise the conclusion
  - ***Example:*** *RNA samples from treatment A were run in one batch (or time 1), and those from treatment B were run in another batch (or time 2).*
- When multiple factors are involved and tested, one-factor design will not work.

Principles of DOE

Four commonly considered principles of DOE (Fisher1935).

- **Representativeness:** Can the experimental units sufficiently represent the conclusion to be made?
- **Randomization:** To avoid unknown or systemic bias
- **Replication:** To increase the precision of the data
- **Error control or blocking:** To reduce known bias (e.g. batch effect).

Experiment needs to be comparative.

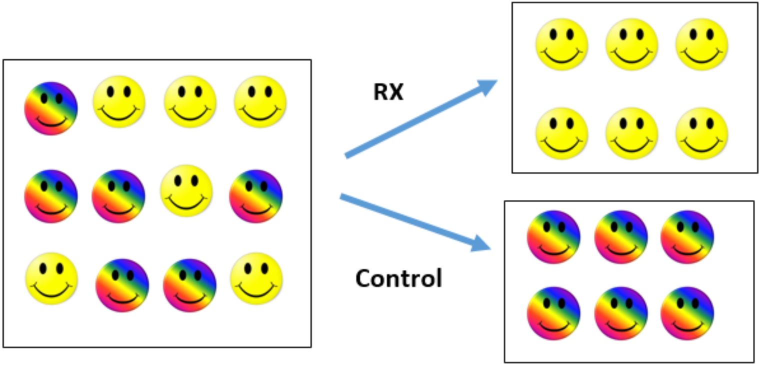
Fisher R.A., 1935 The Design of Experiment, Ed. 2nd Oliver & Boyd, Edingburgh

Representative

"There's a flaw in your experimental design. All the mice are scorpios."

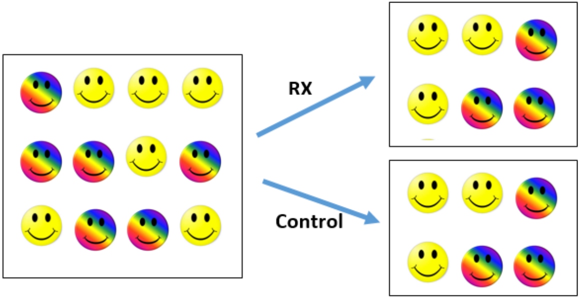
Randomization

Can the following design detect the drug effect?



Randomization

- Each experimental unit should have an equal chance to be assigned to a treatment group or block
- Prevent the introduction of systematic bias into the response of the experiment.
- Allow estimating experimental error.



Replications and Blocking

- **Replications:** Essential for controlling data variation. Why?
  - Observed data: $(Y_1, Y_2, \cdots, Y_n) \sim N(\mu, \sigma^2)$.
  - μ and σ^2 are unknown population parameters.
  - Estimates: $\hat{\mu} = \bar{Y}$ (sample mean) and $\hat{\sigma^2} = S^2$ (sample variance)
  - Standard error of the mean $= \sqrt{S^2/n}$, which determines the confidence interval (CI) of $\hat{\mu}$.
  - larger n (more replications) → narrower CI → more precision in mean estimate.

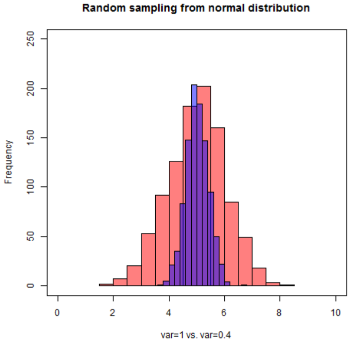
Replications and Blocking

- **Replications:** Essential for controlling data variation. Why?
  - Observed data: $(Y_1, Y_2, \cdots, Y_n) \sim N(\mu, \sigma^2)$.
  - μ and σ^2 are unknown population parameters.
  - Estimates: $\hat{\mu} = \bar{Y}$ (sample mean) and $\hat{\sigma^2} = S^2$ (sample variance)
  - Standard error of the mean $= \sqrt{S^2/n}$, which determines the confidence interval (CI) of $\hat{\mu}$.
  - larger n (more replications) → narrower CI → more precision in mean estimate.
- **Blocking:**
  - Include other factors that contribute to the unwanted variation in the design.
  - By blocking, we can reduce the source of variation.
  - Reduced standard error → narrower CI → more precision in mean estimate.

Accuracy vs. Precision

A well design experiment should generate high quality data.

- **Accuracy:**
    - Focus on if a method or technique produces measurements that are close to the true values.
    - Minimise measurement bias.
    - Microarry vs. RNA-Seq
- **Precision:**
    - Emphasize on smaller variation of the data
    - Lower variation, higher precision because measurements are closer to the mean.



Basic Statistics for DOE

Population and Samples

- **Population:** All possible items, units, or subjects from an experimental or observational condition.
- **Samples:** A group of units taken from a population.
- **Statistics uses samples to make inferences about the entire population.**

Example:

- All cancer patients in the Duke hospital vs. patients consented to participate in a research study.
- Tumor vs. tumor cells extracted for an experiment

Random variable

- **Random variable (Y):** A variable represents all possible observations (measurements) collected for a study
  - Quantitative: continuous measures
  - Qualitative: binary, categorical, counts
- Assuming observed continuous data $y_i, i = 1, \cdots, n$

$$y_i = \mu + \epsilon_i, i = 1, \cdots, n$$

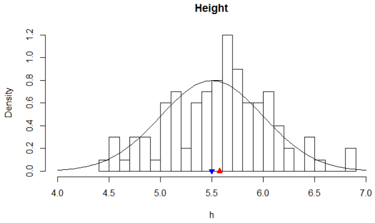
  - μ: unknown population parameter of interest.
  - ϵ: random and unobserved variable; $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$ are independent and follow a normal distribution $N(0, \sigma^2)$.
  - $Var(\epsilon) = \sigma^2 = Var(Y)$, an unknown population parameter

Illustration

For a random variable Y, y_i is the i^{th} observed value,$i = 1, \cdots, n$

- **Sample mean** $\bar{y} = \frac{\sum_i^n y_i}{n}$
- **Sample variance** $S^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$

Example: Assume the true distribution of the height of high school Seniors is a normal distribution $N(\mu = 5.5, \sigma^2 = 0.25)$. We randomly survey 100 students for their height.



Average height, $\bar{y} = 5.57$
Sample variance, $S^2 = 0.2495$

Example: height of the high school Seniors

If we survey 20, 100, and 500 students, can we make a good inference for the student height?

- Assume 10,000 random samples from $N(5.5, 0.25)$ as the 'population' of the high school students.
- Randomly draw 20, 100, and 500 values from the population (10,000 data points).

Sample size, n	20	100	500
Sample Mean	5.458	5.509	**5.493**
Sample Variance	0.297	0.191	**0.241**

Example: height of the high school Seniors

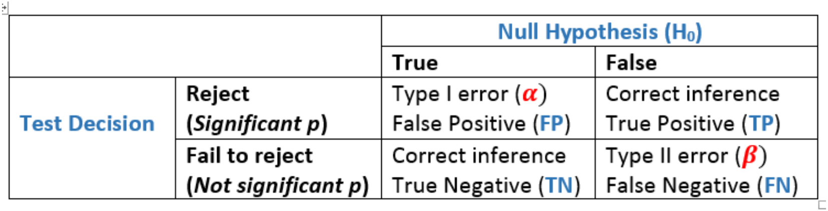
If we survey 20, 100, and 500 students, can we make a good inference for the student height?

- Assume 10,000 random samples from $N(5.5, 0.25)$ as the 'population' of the high school students.
- Randomly draw 20, 100, and 500 values from the population (10,000 data points).

Sample size, n	20	100	500
Sample Mean	5.458	5.509	**5.493**
Sample Variance	0.297	0.191	**0.241**

- **Random variation can have a bigger effect on sample estimates in small group. Sample size matters**
- Critical for precision of estimates
- Critical for statistical power in hypothesis testing

Statistical power

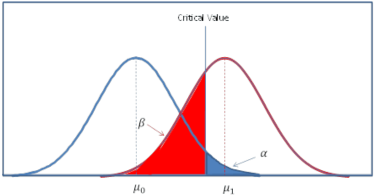


		Null Hypothesis (H_0)	
		True	**False**
Test Decision	**Reject** **(*Significant p*)**	Type I error (***α***) False Positive (**FP**)	Correct inference True Positive (**TP**)
	Fail to reject **(*Not significant p*)**	Correct inference True Negative (**TN**)	Type II error (***β***) False Negative (**FN**)

$$\text{Power} = 1 - \beta$$

Power and Sample Size

A well-designed study should have sufficient statistical power.



- Determine what test statistics to be used for the hypothesis testing.
- Assume a two-sample t-test, the effect size is

$$\Delta = \frac{|\mu_0 - \mu_1|}{\sigma}$$

- The sample size is $n = 2\frac{(Z_{1-\alpha}+Z_{1-\beta})^2}{\Delta^2}$
- The larger the effect size, the smaller n.

Key elements for power calculation: (1) study design; (2) statistical methods; (3) some ideas of target 'effect size' from literature or pilot study.

Considerations behind analysis methods

- Experimental or study design.
- Types of the response (dependent) variable:
  - continuous or discrete data; distribution of the data
  - binary or categorical
- Types of predictor variable: continuous vs. categorical
- Any covariates to be adjusted?

Example: The two-factor RNA-Seq experiment in this workshop

- Dependent variable: Gene expression
- Factors: strain, media
- Any covariates?

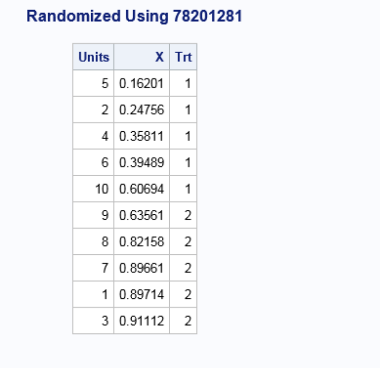
Types of Designs

Completely Randomized Design (CRD)

- Assume homogeneous experimental units.
- Factor considered is 'categorical'. It can be two or multiple levels/groups.
  Example: Treatment groups (YPD vs. TC)
- **Randomization:** Each experimental unit has an equal likely chance to be assigned to a treatment group. Assume *t* treatment groups and *n* experimental units per group, totally *nt* experimental units.
  1. Label experimental units 1 to *nt*.
  2. Generate a random number for each experimental unit (keep the label and random number paired).
  3. Rank the random number, and the first *n* units go to treatment 1, 2nd set of *n* units go to treatment 2, etc.

Example: Plan to randomly assign two different growth media (YPD and TC) to 10 H99 strain before RNA extraction.

- Designate sample ID number 1 to 10.
- Use a seed number (e.g. 78201281) to generate 10 random numbers (x) between 0 and 1 for each sample.
- Sort x from low to high
- Assign the first 5 to treatment 1 and the rest to treatment 2.



Randomized Using 78201281

Units	X	Trt
5	0.16201	1
2	0.24756	1
4	0.35811	1
6	0.39489	1
10	0.60694	1
9	0.63561	2
8	0.82158	2
7	0.89661	2
1	0.89714	2
3	0.91112	2

Measurements of variation

1. Assume observation units are continuous measurements
2. n samples obtained from each treatment group:
   Within group variation: $S^2 = \frac{\sum_i^n (y_i - \bar{y})^2}{n-1}$
3. t treatment groups, n samples per group:
   Between treatment variation:

$$MST = \frac{n \sum_i^t (\bar{y_i}. - \bar{y})^2}{t-1}$$

Within treatment variation:

$$MSE = \frac{\sum_i^t \sum_j^n (y_{ij} - \bar{y_i}.)^2}{t(n-1)}$$

Data analysis for CRD

Dependent variable: Gene expression level (y_{ij})
Independent variable: Treatment group (β_i)

Model: $y_{ij} = \mu + \beta_i + \epsilon_{ij}$, $i = 1, \cdots, t$ and $j = 1, \cdots, n$

Analysis of variance (ANOVA)Table:

Source	df	Mean SS (MS)	F
Treatment	$t-1$	MST	$\frac{MST}{MSE}$
Error	t(n-1)	MSE	

$F = \frac{\text{Variation between treatments}}{\text{Variation within treatment}}$,

following an F distribution with d.f. of $(t-1, t(n-1))$.

one-way ANOVA example: PlantGrowth

```
plant <- PlantGrowth
plant
```

```
##    weight group
## 1    4.17  ctrl
## 2    5.58  ctrl
## 3    5.18  ctrl
## 4    6.11  ctrl
## 5    4.50  ctrl
## 6    4.61  ctrl
## 7    5.17  ctrl
## 8    4.53  ctrl
## 9    5.33  ctrl
## 10   5.14  ctrl
## 11   4.81  trt1
## 12   4.17  trt1
## 13   4.41  trt1
## 14   3.59  trt1
## 15   5.87  trt1
## 16   3.83  trt1
## 17   6.03  trt1
## 18   4.89  trt1
## 19   4.32  trt1
## 20   4.69  trt1
## 21   6.31  trt2
## 22   5.12  trt2
## 23   5.54  trt2
## 24   5.50  trt2
## 25   5.37  trt2
## 26   5.29  trt2
## 27   4.92  trt2
```

PlantGrowth dataset in R for plant yield (dried weight of plants) of 30 plants, which were randomized to three treatment groups (control, treatment 1, treatment 2).

```
res <- anova(lm(plant$weight ~ plant$group, data = plant))
res <- data.frame(res)
res
```

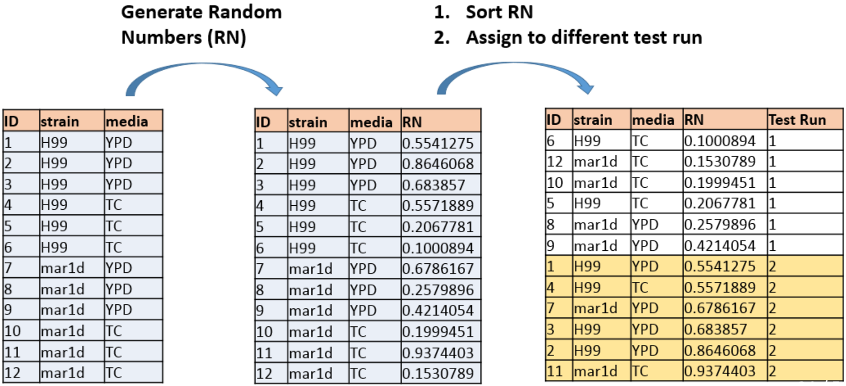
```
##             Df    Sum.Sq   Mean.Sq  F.value      Pr..F.
## plant$group  2   3.76634 1.8831700 4.846088 0.01590996
## Residuals   27  10.49209 0.3885959       NA          NA
```

CRD Pros and Cons

- **Pros:**
  - Easy to randomize experimental units
  - Simple statistical analysis: two sample t-test, one-way ANOVA, generalized linear regression if data is not normal distributed (e.g. negative binomial for RNA-Seq read counts)
  - Flexible in terms of number of experimental units per groups (equal or unequal number per group).
- **Cons:** Can't control the differences between experimental units prior to the randomization.
  Example: If there are more females than males in the study,
  - CRD cannot control the gender effect.
- For CRD, it is better to have homogeneous experimental units or large sample size.

Factorial experiments in CRD

- A factorial experiment includes all possible factor-level combinations in the experiment, for instance, strain-media combinations and their replicates.
- Follow CRD to group samples for different experiment runs (test run)



Generate Random Numbers (RN)

1. **Sort RN**
2. **Assign to different test run**

ID	strain	media
1	H99	YPD
2	H99	YPD
3	H99	YPD
4	H99	TC
5	H99	TC
6	H99	TC
7	mar1d	YPD
8	mar1d	YPD
9	mar1d	YPD
10	mar1d	TC
11	mar1d	TC
12	mar1d	TC

ID	strain	media	RN
1	H99	YPD	0.5541275
2	H99	YPD	0.8646068
3	H99	YPD	0.683857
4	H99	TC	0.5571889
5	H99	TC	0.2067781
6	H99	TC	0.1000894
7	mar1d	YPD	0.6786167
8	mar1d	YPD	0.2579896
9	mar1d	YPD	0.4214054
10	mar1d	TC	0.1999451
11	mar1d	TC	0.9374403
12	mar1d	TC	0.1530789

ID	strain	media	RN	Test Run
6	H99	TC	0.1000894	1
12	mar1d	TC	0.1530789	1
10	mar1d	TC	0.1999451	1
5	H99	TC	0.2067781	1
8	mar1d	YPD	0.2579896	1
9	mar1d	YPD	0.4214054	1
1	H99	YPD	0.5541275	2
4	H99	TC	0.5571889	2
7	mar1d	YPD	0.6786167	2
3	H99	YPD	0.683857	2
2	H99	YPD	0.8646068	2
11	mar1d	TC	0.9374403	2

Randomized Completed Block Design(RCBD)

- Probably most frequently used design
- **Goal**: Minimize the effect of nuisance factors to the observation units.
- **Types of nuisance factors**: different technicians, different days(time) of experiment, etc.
- Restrict randomization to homogeneous blocks.
- Block is usually treated as a random effect.

How the RCBD works?

- Identify nuisance factor to be used for blocking.
- Sort experimental units into homogeneous batches (blocks). The experimental units within each batch is as uniform as possible.
- Proceed with CRD within each block: randomly assign treatments to experiments units within each block.
- **Model:** Factors to considered: blocks (β_i), treatments (τ_j). ANOVA model:

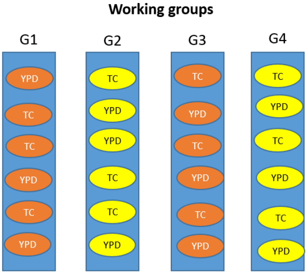
$$y_{ijk} = \mu + \beta_i + \tau_j + \epsilon_{ijk},$$

where $i = 1, \cdots, b$ for blocks, $j = 1, \cdots, t$ for treatments, $k = 1, \cdots, k$ for replicates in each treatment-block combination, and $\epsilon_{ijk} \sim N(0, \sigma^2)$

Illustration

Four working group will complete an experiment of 24 samples (6 samples per strain×media combination). Each group will handle 6 samples (3 per media group). We can consider each working group as a homogeneous block.

- Randomly assign 6 samples of the same strain (H99 or mar1d) to each working group (*i.e.* 6 samples per block).
- Randomly assign two treatments (YPD and TC) to samples handled by each working group (within each block).



Two-way ANOVA example: Stress reduction example

```
stress <- read.csv(file = "./data/stress.csv")
stress <- data.frame(stress)
stress


##    Treatment   Age StressReduction
## 1     mental young              10
## 2     mental young               9
## 3     mental young               8
## 4     mental   mid               7
## 5     mental   mid               6
## 6     mental   mid               5
## 7     mental   old               4
## 8     mental   old               3
## 9     mental   old               2
## 10  physical young               9
## 11  physical young               8
## 12  physical young               7
## 13  physical   mid               6
## 14  physical   mid               5
## 15  physical   mid               4
## 16  physical   old               3
## 17  physical   old               2
## 18  physical   old               1
## 19   medical young               8
## 20   medical young               7
## 21   medical young               6
## 22   medical   mid               5
## 23   medical   mid               4
## 24   medical   mid               3
## 25   medical   old               2
## 26   medical   old               1
```

27 subjects from three age groups (young, mid, and old ages) were studied for stress reduction by three types of stress reduction treatments (mental, physical, and medical).

```
res <- anova(lm(StressReduction ~ Treatment + Age, data = stress))
res <- data.frame(res)
res


##           Df Sum.Sq    Mean.Sq F.value       Pr..F.
## Treatment  2     18  9.0000000      11 4.882812e-04
## Age        2    162 81.0000000      99 1.000000e-11
## Residuals 22     18  0.8181818      NA           NA
```

In this example, $b = 3$ for age groups, $t = 3$ for treatment groups, and $k = 3$ for repeats within each block-treatment combination.

RCBD Pros and Cons

- **Pros:**
  - Good for comparing treatment effect when there is one nuisance factor to worry about.
  - Easy to construct the experiment
  - Simple statistical analysis
  - Flexible for any numbers of treatments and blocks.
- **Cons:**
  - It can only control variability from one nuisance factor.
  - Since it requires homogeneous blocks, it is better for a study with a small number of treatments (factor levels) to test.
  - It requires the number of experimental units $\geq$ the number of factor-level combinations of interest.

RNA-Seq Design

Designs for RNA-Seq experiment

Reference paper: Auer and Doerge, Genetics, 2010

Statistical Design and Analysis of RNA Sequencing Data

Paul L. Auer and R. W. Doerge[1]

Department of Statistics, Purdue University, West Lafayette, Indiana 47907

Manuscript received January 31, 2010
Accepted for publication March 15, 2010

RNA-Seq Experiment

Steps of a RNA-Seq experiment

1. RNA is isolated from cells, fragmented at random positions, and copied into complementary DNA (cDNA)
2. Fragments meeting a certain specified size (*e.g.* 200 – 300 bp) are retained for PCR
3. Sequencing
4. Sequence alignment to generate sequence reads at each position
5. **Data:** Counts of sequence reads or **digital gene expression (DGE)**
6. **Types of reads:** junction reads, exonic reads, polyA reads

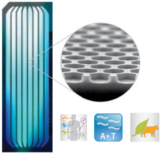
Sources of variability

1. Biological variability
   - Variability between experimental units (samples)
   - Variability between factors of interest (treatment groups)
   - Biological variability is not affected by technical variability.

2. Technical variability:

- between sequencing platforms
- between library construction
- between flow cells (different runs)
- between lanes

Flow cells: A glass slide with 1, 2, or 8 separate lanes (Illumina RNA-Seq)

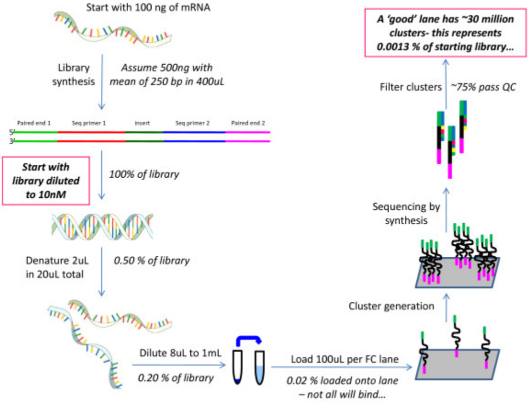


References: Marioni et al. Genome Res. 2008; McIntyre et al. BMC Genomics 2011

Sampling in RNA-Seq

- **Subject sampling:** Subjects (*e.g.* organisms or individuals) are ideally drawn from a large population to which the results can be generalized.
- **RNA sampling:** occurs during the experimental procedure when RNA is isolated from the cell(s).
- **Fragment sampling:** Only certain fragmented RNAs are retained for amplification. The sequencing reads do not represent 100% of the fragments loaded into a flow cell resulted in fragment sampling.

More on RNA and fragment sampling



Start with 100 ng of mRNA

Library synthesis

Assume 500ng with mean of 250 bp in 400uL

Start with library diluted to 10nM

100% of library

Denature 2uL in 20uL total

0.50 % of library

Dilute 8uL to 1mL

0.20 % of library

Load 100uL per FC lane

0.02 % loaded onto lane – not all will bind...

Cluster generation

Sequencing by synthesis

Filter clusters

~75% pass QC

A 'good' lane has ~30 million clusters- this represents 0.0013 % of starting library...

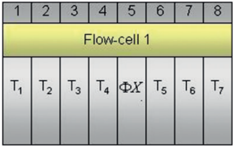
Library concentration 10nM=4pM → $\frac{4}{10^{12}} \times 6.02 \times 10^{23} = 2.408 \times 10^{12}$ total molecules in the library

→ $\frac{30{,}000{,}000}{2.408 \times 10^{12}} = 0.0013\%$ of molecules to be analyzed. (McIntyre et al. 2011)

Unreplicated data

Outline of experiment:

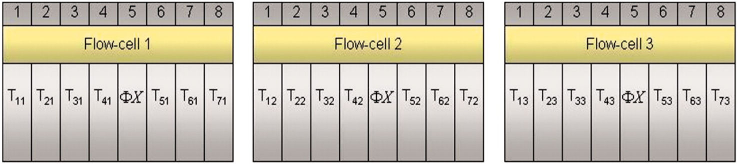
- mRNA isolated from subjects within different treatment group ($T_1, \cdots, T_7$).
- a ΦX genomic sample is loaded to lane 5 as a control
- ΦX can be used to recalibrate the quality score of sequencing reads from other lane.



Problems:

- Lack of knowledge about biological variation
- Unable to estimate within treatment variation leading to no basis for inference of between treatment effect.
- Results are specific to the subjects in the study and can't be generalized.

Replicated data: Multiple flow-cell design



1	2	3	4	5	6	7	8
Flow-cell 1							
T_{11}	T_{21}	T_{31}	T_{41}	ΦX	T_{51}	T_{61}	T_{71}

1	2	3	4	5	6	7	8
Flow-cell 2							
T_{12}	T_{22}	T_{32}	T_{42}	ΦX	T_{52}	T_{62}	T_{72}

1	2	3	4	5	6	7	8
Flow-cell 3							
T_{13}	T_{23}	T_{33}	T_{43}	ΦX	T_{53}	T_{63}	T_{73}

- **Exp Design:** Seven treatment groups, three biological replicates, and one sample per lane. T_{ij} for i^{th} treatment group and j^{th} replicate. $i = 1, \cdots, 7$ and $j = 1 - 3$.
- **Factor of consideration:** treatment effect (τ_{ik}) for gene k.

$$(\text{Dependent variable})_{ijk} = \alpha_k + \tau_{ik} + \epsilon_{ijk}$$

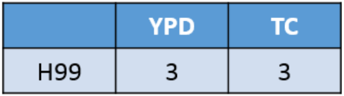
- **Problem:** Cannot separate treatment effect from technical effect since biological replicates are run in different flow-cells.

Balanced block design

- **Objective:** To control two sources of technical variation: batch effect and lane effect.
- **Multiplexing:** All samples are pooled to be run within the same lane.
    - Take the advantage of bar coding of RNA fragments.
    - To keep the same sequence depth, divide the amplification product to run in multiple lanes
    - If (# of lanes) = (# of samples), it produces the same sequence depth as running one sample per lane.
    - Each lane has the same set of samples – eliminate the lane effect

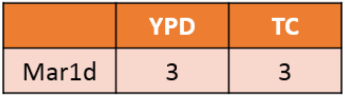
How will you randomize samples in your experiment?

RNA-Seq for samples from two Cryptococcus neoformans strains under two growth media



	YPD	TC
H99	3	3

or



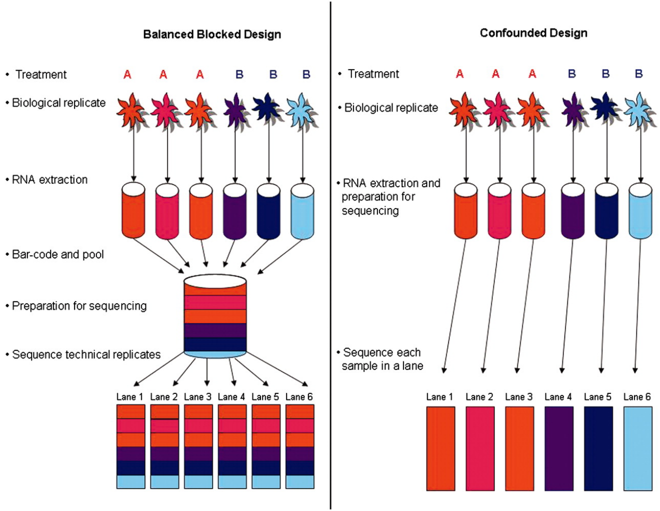
	YPD	TC
Mar1d	3	3

1. Each working group has 6 samples, 3 per treatment group
2. Four working groups to complete 6 samples per strain×media (24 samples total).
3. Another four working groups to repeat the same set of samples.

Balanced Block Design - I (BBD I)

- Three biological replicates per treatment (growth media) ($j = 1, \cdots, 3$)
- Two growth meadia (YPD and TC) ($i = 1, \cdots, 2$)
- RNA are bar-coded and pooled
- Divide the pool to six equal subset to run on 6 lanes (six technical replicates, $t = 1, \cdots, 6$)
- Single flow cell run

BBD vs. Confounded design



Analysis model for BBD I

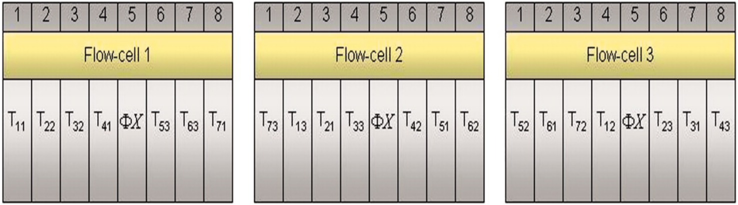
- **Dependent variable:** DGE measures, defined by the distribution you assumed for the sequence reads. For example,
  - Auer et al. assumed $y_{ijk} \sim Possion(\mu_{ijk})$.
  - DESeq2 uses Negative Binomial model.

  $y_{ijk} = \sum_t y_{ijkt}$, where i for treatment, j for sample, k for gene, and t for the 6 technical replicates
- **Factors considered in the GLM:** treatment effect (τ_{ik}) since all samples are from a single strain.

$$(\text{Dependent variable})_{ijk} = \alpha_k + \tau_{ik} + \epsilon_{ijk}$$

- No lane effect was included in this model as they considered lane effects were balanced across treatment groups.
- No batch effect in this case since it is only one flow-cell run.
- Each working group can analyze their own data.

Balanced block design II (BBD II) - without multiplexing



	1	2	3	4	5	6	7	8
Flow-cell 1	T_{11}	T_{22}	T_{32}	T_{41}	ΦX	T_{53}	T_{63}	T_{71}
Flow-cell 2	T_{73}	T_{13}	T_{21}	T_{33}	ΦX	T_{42}	T_{51}	T_{62}
Flow-cell 3	T_{52}	T_{61}	T_{72}	T_{12}	ΦX	T_{23}	T_{31}	T_{43}

- A design that can run one sample per lane but also has good randomization of samples within each flow-cell.
- Three biological replicates within seven treatment groups. T_{ij}, where $i = 1, \cdots, 7$ for treatment groups and $j = 1, \cdots, 3$ for samples.
- **Two block effects:** flow cells and lanes.

Analysis for BBD II

- **Dependent variable:** Same as before, but it is coded to indicate treatment (i), flow-cell (f), lane (l), and gene (k).
- **Factors to consider:** treatment effect (τ_{ik}), flow-cell effect (ν_{fk}), and lane effect (ω_{lk}).

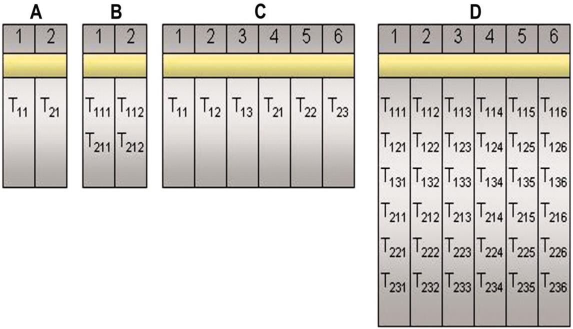
$$\text{(Dependent variable)}_{ijflk} = \alpha_k + \tau_{ik} + \nu_{fk} + \omega_{lk} + \epsilon_{ijflk}$$

ϵ_{ijflk} is the error term.

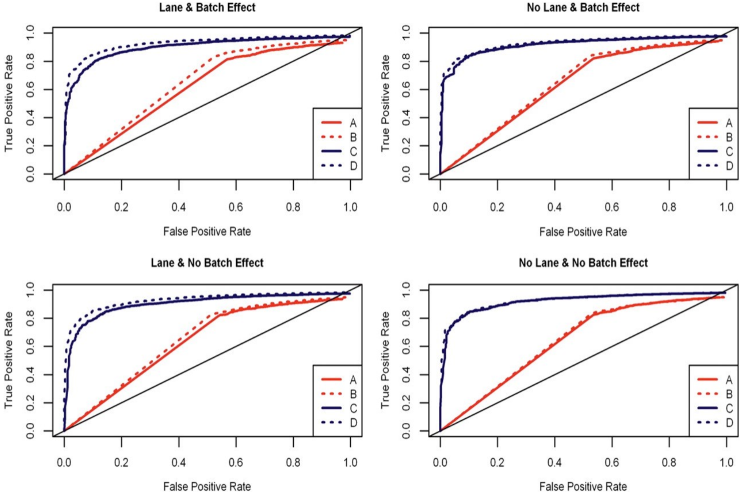
Summary for Balanced block design

- The feature of unique bar-code for RNA fragments in RNA-Seq makes blocking design possible.
- Can control batch and lane effects
- Multiplex design illustrated here requires the number of unique bar-codes equal or greater than the samples in each lane.
- For Illumina, a total of 12 unique barcodes can be used in one lane. Therefore, 96 samples can be multiplexed in one flow-cell run.

Performance comparison between designs by simulation studies


A
1 2
T_{11} T_{21}
B
1 2
T_{111} T_{112}
T_{211} T_{212}
C
1 2 3 4 5 6
T_{11} T_{12} T_{13} T_{21} T_{22} T_{23}
D
1 2 3 4 5 6
T_{111} T_{112} T_{113} T_{114} T_{115} T_{116}
T_{121} T_{122} T_{123} T_{124} T_{125} T_{126}
T_{131} T_{132} T_{133} T_{134} T_{135} T_{136}
T_{211} T_{212} T_{213} T_{214} T_{215} T_{216}
T_{221} T_{222} T_{223} T_{224} T_{225} T_{226}
T_{231} T_{232} T_{233} T_{234} T_{235} T_{236}

T_{ijk}: i for treatment, j for sample, k for technical replicates.
A: unreplicated data; **B:** no biological replicates, two technical replicates (BBD without biological replicates); **C:** no technical replicates (unblocked design); **D**:BBD with biological and technical replicates.



C&D always perform better than A&B. When simulation included lane and/or batch effects, **D (balanced block design)** performed better than **C (unblocked design)**.

Summary

- Outline a testable hypothesis.
- Identify factor(s) of interest and nuisance factors to be controlled and then determine the type of experimental design to use.
- Follow the four key principles of DOE. These classical principles still apply to RNA-Seq.
- Statistical model should reflect to the experimental design.
- Sample size should be determined based on power calculation prior to the study.
- Technical variation exists and should be taken into account in RNA-Seq.
  - Lane effect, batch effect
- Multiplexing in NGS allow us to implement randomization and blocking.

References

- Marioni et al. Genome Res. (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays
- McIntyre et al. BMC Genomics (2011) RNA-seq: technical variability and sampling
- Auer and Doerge Genetics (2010) Statistical Design and Analysis of RNA Sequencing Data
- Planning, Construction, and Statistical Analysis of Comparative Experiments, Francis G. Giesbrecht and Marcia L. Gumpertz (Wiley)