Statistical Analysis of RNA-Seq Data: Experimental design

Lorena S. Rivarola-Duarte

PhD Student

HELMHOLTZ CENTRE FOR ENVIRONMENTAL RESEARCH – UFZ

UNIVERSITÄT LEIPZIG

Introduction

- Next Generation Sequencing (NGS) is becoming the **preferred approach** for characterizing and quantifying transcriptomes.
- Even though the data produced is really informative, little attention has been paid to **fundamental design aspects** of data collection:
  - Sampling
  - Randomization
  - Replication
  - Blocking

Introduction

- Next Generation Sequencing (NGS) is becoming the **preferred approach** for characterizing and quantifying transcriptomes.
- Even though the data produced is really informative, little attention has been paid to **fundamental design aspects** of data collection:
  - Sampling
  - Randomization
  - Replication
  - Blocking

Discussion of these concepts in an RNA-seq framework

Introduction

RNA-seq uses NGS technology (Illumina, 454, SOLiD) to sequence, map and quantify a population of transcripts

Advantages

- Greater sensitivity than microarrays,
- Able to discriminate closely homologous regions,
- Does not require a priori assumptions about regions of expression.

There are many steps in the experimental process that may introduce **errors and biases**

Introduction

RNA-seq uses NGS technology (Illumina, 454, SOLiD) to sequence, map and quantify a population of transcripts

Advantages

- Greater sensitivity than microarrays,
- Able to discriminate closely homologous regions,
- Does not require a priori assumptions about regions of expression.

There are many steps in the experimental process that may introduce **errors and biases**

Introduction

RNA-seq uses NGS technology (Illumina, 454, SOLiD) to sequence, map and quantify a population of transcripts

Advantages

- Greater sensitivity than microarrays,
- Able to discriminate closely homologous regions,
- Does not require a priori assumptions about regions of expression.

There are many steps in the experimental process that may introduce **errors and biases**

Methodology:

- RNA is isolated from cells,
- Fragmented at random positions,
- Copied into complementary DNA,
- Selection of fragments with a certain size range,
- Amplification using PCR,
- Sequencing,
- Reads are aligned to a reference genome,
- The number of sequencing reads mapped to each gene in the reference is tabulated.

Methodology:

- RNA is isolated from cells,
- Fragmented at random positions,
- Copied into complementary DNA,
- Selection of fragments with a certain size range,
- Amplification using PCR,
- Sequencing,
- Reads are aligned to a reference genome,
- The number of sequencing reads mapped to each gene in the reference is tabulated.

These gene counts or **digital gene expression (DGE)** can be used to test differential gene expression

Introduction

- Soon after the introduction of microarrays researchers discuss about the **need for proper experimental design** (Keer *et al*, 2000), and the application of the fundamental aspects formalized by **Fisher** in 1935.
- Randomization – Replication – Blocking

Introduction

- Soon after the introduction of microarrays researchers discuss about the **need for proper experimental design** (Keer *et al*, 2000), and the application of the fundamental aspects formalized by **Fisher** in 1935.
- Randomization – Replication – Blocking

Now we need the same for RNA-seq data!

Experimental Design

- The experimenter is often interested in the **effect** of some process or intervention (the "treatment") on some **objects** (the "experimental units").
- For differential expression analyses, researchers are interested in comparisons across treatment groups in the form of contrasts or pairwise comparisons.

Experimental Design

Randomization

It is the process of assigning individuals at random to groups or to different groups in an experiment.

This **reduces bias** by equalising so-called factors (independent variables) that have not been accounted for in the experimental design.

Experimental Design

Replication

Measurements are usually subject to variation and uncertainty.

Then, measurements are repeated and full experiments are replicated to help **identify the sources of variation**, to better estimate the true effects of treatments, to further strengthen the experiment's reliability and validity.

Experimental Design

Blocking

Experimental units are grouped into homogeneous clusters in an attempt to improve the comparison of treatments by randomly allocating the treatments within each cluster or 'block'.

Blocking **reduces** known but **irrelevant sources of variation** between units and thus allows greater precision in the estimation of the source of variation under study.

Experimental Design

Example

Effectiveness of 2 different diets.

Many different subjects (replication) recruited from multiple weight loss centers (blocking) and each center would randomly assign its subjects to one of two diets (randomization).

Experimental Design

- These principles are well known but their **implementation** often requires significant planning and statistical expertise.
- In the absence of a proper design, it is impossible to partition biological variation from technical variation.
- No amount of statistical sophistication can separate **confounded factors** AFTER data have been collected

RNA-seq: Sampling

Regardless of the design, we have 3 levels of sampling:

- **Subject sampling:** individuals are ideally drawn from a larger population to which results of the study may be generalized.
- **RNA sampling:** occurs during the experimental procedure when RNA is isolated from the cell.
- **Fragment sampling:** only certain fragmented RNAs that are sampled from the cells are retained for amplification. Since the sequencing reads do not represent 100% of the fragments loaded into a flow cells, this is also at play.

RNA-seq: Sampling

Regardless of the design, we have 3 levels of sampling:

- **Subject sampling:** individuals are ideally drawn from a larger population to which results of the study may be generalized.
- **RNA sampling:** occurs during the experimental procedure when RNA is isolated from the cell.
- **Fragment sampling:** only certain fragmented RNAs that are sampled from the cells are retained for amplification. Since the sequencing reads do not represent 100% of the fragments loaded into a flow cells, this is also at play.

RNA-seq: Sampling

Regardless of the design, we have 3 levels of sampling:

- **Subject sampling:** individuals are ideally drawn from a larger population to which results of the study may be generalized.
- **RNA sampling:** occurs during the experimental procedure when RNA is isolated from the cell.
- **Fragment sampling:** only certain fragmented RNAs that are sampled from the cells are retained for amplification. Since the sequencing reads do not represent 100% of the fragments loaded into a flow cells, this is also at play.

Library complexity!!!

RNA-seq: library complexity

How to achieve a high complexity library, or normalized RNA-seq libraries?

- Crab duplex-specific nuclease: When double stranded cDNA is denatured and then allowed to partially re-anneal, the most abundant species –which re-anneal quicker- are digested with a nuclease, decreasing the proportion of these reads 50x and enrich the lower-expressed 10x. (Christodoulou *et al* 2011)

- "Comprehensive comparative analysis of strand-specific RNA sequencing methods". Levin *et al* 2010. Nature Methods, 7:9.

RNA-seq: library complexity

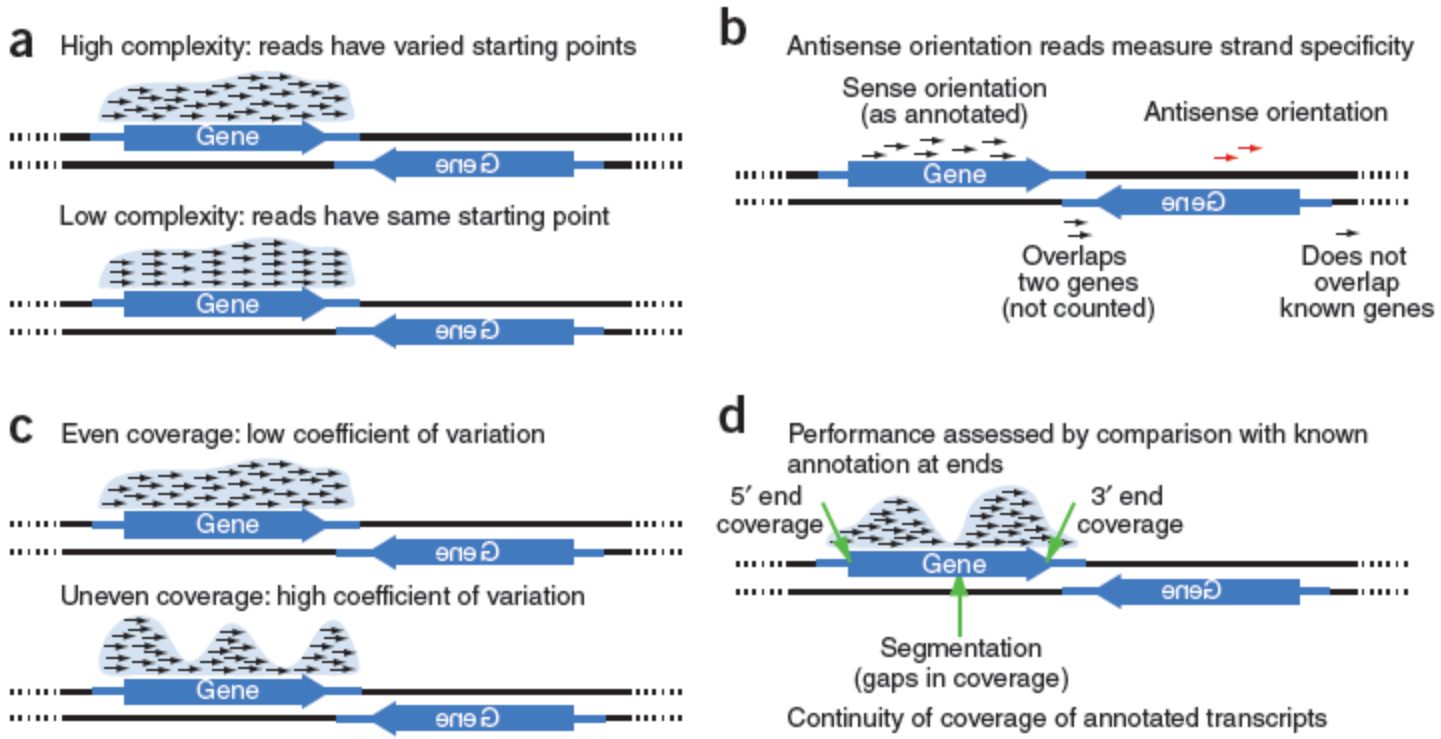


Figure 2 | Key criteria for evaluation of strand-specific RNA-seq libraries. (**a**–**d**) Categories of quality assessment were complexity (**a**), strand specificity (**b**), evenness of coverage (**c**) and comparison to known transcript structure (**d**). Double-stranded genome with gene ORF orientation (blue arrows) and UTRs (blue lines) are shown along with mapped reads (black and red arrows, reads mapped to sense and antisense strands, respectively).

RNA-seq: Unreplicated data

- Observational studies with **no biological replication**.
- The assigment of subjects to treatment groups is not decided by the investigator.
- Example: mRNA isolated from liver and kidney tissues (extracted from one human cadaver) randomly fragmented and sequenced. The different treatments consist of different tissues.

RNA-seq: Unreplicated data

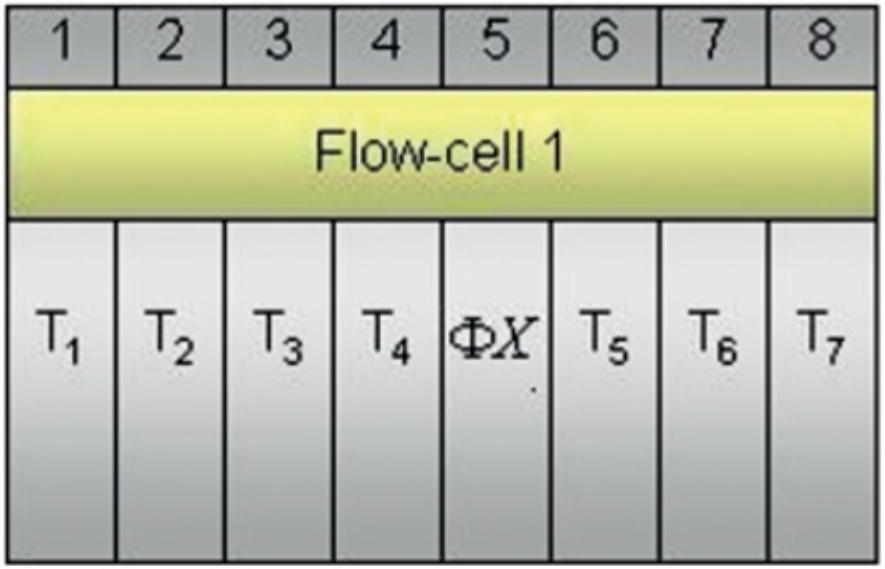


FIGURE 1.—Hypothetical Illumina GA flow cell with mRNA isolated from subjects within seven different treatment groups ($T_1, \ldots, T_7$) and loaded into individual lanes (*e.g.*, the mRNA from the subject within treatment group 1 is sequenced in lane 1). As a control, a ΦX genomic sample is often loaded into lane 5. The bacteriophage ΦX genome is known exactly and can be used to recalibrate the quality scoring of sequencing reads from other lanes (BENTLEY *et al.* 2008).

RNA-seq: Unreplicated data

- Data analysis proceeds on a gene by gene basis organizing the data in a **2x2 table**.
- **Fisher´s exact test:**
  - Used in the analysis of contingency tables.
  - The significance of the deviation from a null hypothesis can be calculated exactly, rather than relying on an approximation that becomes exact in the limit as the sample size grows to infinity.

TABLE 1

A 2 × 2 contingency table of (unreplicated) digital gene expression (DGE) measures for testing differential expression between Treatment$_1$ and Treatment$_2$ of gene A

	Treatment$_1$	Treatment$_2$	Total
Gene A	n_{11}	n_{12}	$N_{1.}$
Remaining genes	n_{21}	n_{22}	$N_{2.}$
Total	$N_{.1}$	$N_{.2}$	$N_{..}$

The cell counts n_{ki} represent the DGE count for gene A ($k = 1$) or the remaining genes ($k = 2$) for Treatment$_i$, $I = 1, 2$. The kth marginal row total is denoted $N_{k.}$, $N_{.i}$ is the marginal total for column i, and $N_{..}$ is the grand total.

Which is the probability of observing an outcome at least as unlikely as n_{11} gene A? if this probability is small then the column classification (treatment) has affected the gene expression

RNA-seq: Unreplicated data

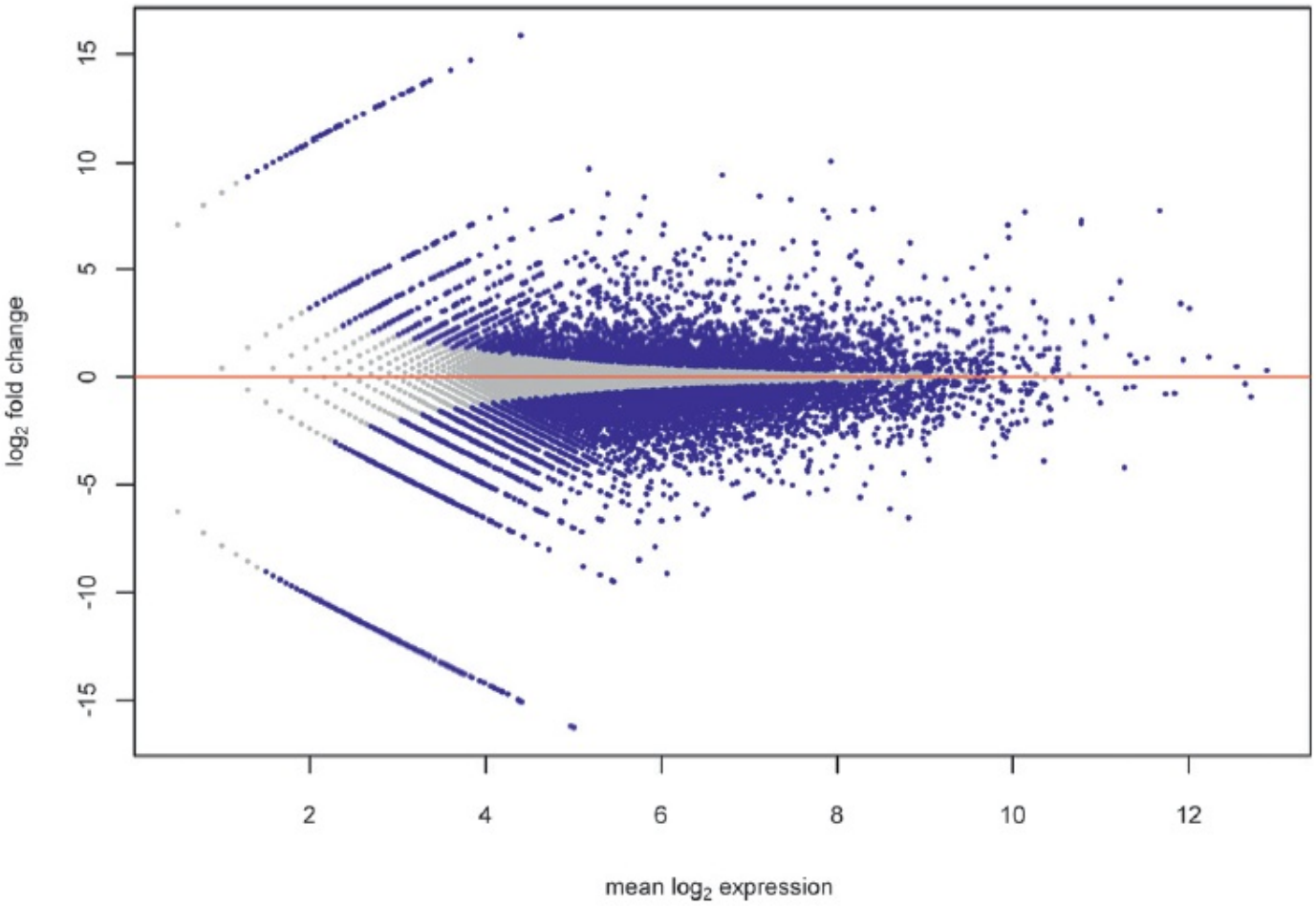


FIGURE 2.—The $\log_2$ fold change, between Treatment$_1$ and Treatment$_2$, of the normalized gene expression is plotted on the *y*-axis, and the mean $\log_2$ expression is plotted on the *x*-axis. Gene expression counts were normalized by the column totals of the corresponding 2 × 2 table (*e.g.*, Table 1). Blue dots represent significantly differentially expressed genes as established by Fisher's exact test; gray dots represent genes with similar expression. The red horizontal line at zero provides a visual check for symmetry.

- Behavior of Fisher´s exact test for testing differential expression between 2 treatments for every gene in a RNA-seq data set.

RNA-seq: Unreplicated data

Limitations of unreplicated data:

- Complete lack of knowledge about biological variation.
- Without an estimate of variability (i.e. within treatment groups), there is **no basis for inference** (between treatment groups).
- The results only apply to the specific subjects included in the study.

RNA-seq: Replicated data

The biological replicates allow for the estimation of within-treatment group (biological) variability, provide information that is necessary for making inferences between treatment groups, and give rise to **conclusion that can be generalized**.

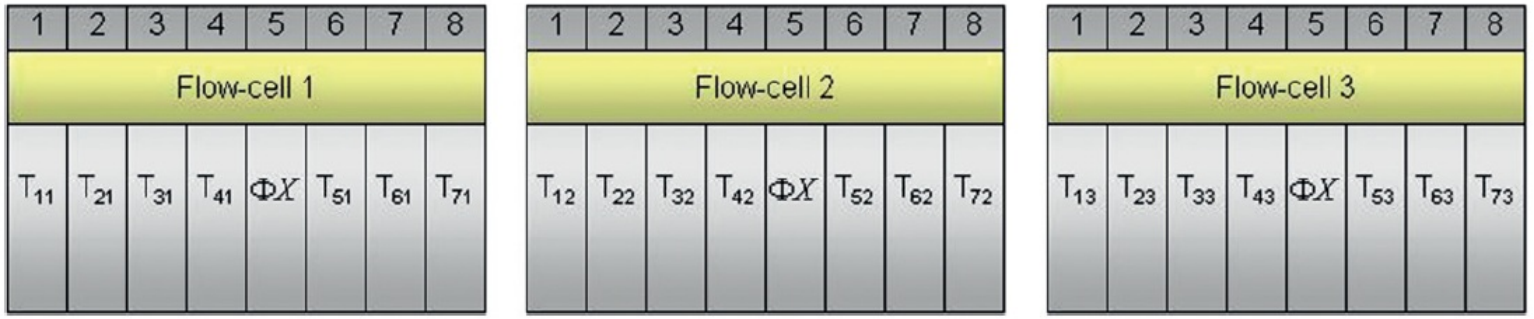


FIGURE 3.—A multiple flow-cell design based on three biological replicates within seven treatment groups. There are three flow cells with eight lanes per flow cell. The control ΦX sample is in lane 5 of each flow cell. T_{ij} refers to the jth replicate in the ith treatment group ($i = 1, \ldots, 7; j = 1, \ldots, 3$).

RNA-seq: Replicated data

- A simple method for testing differential expression that incorporates within-group (or within treatment) variability relies on a **Generalized Linear Model (GLM)** with overdispersion -*var(Y)>E(Y)*-.
- It is a flexible generalization of ordinary **linear regression** that allows for response variables that have **other than a normal distribution**. (Normal distribution is one of the assumptions underlying linear regression)
- When data is counts of events (or items) then a **discrete distribution** (like Poisson) is more appropriate than approximating with a continuous distribution (Negative counts do not make sense).

RNA-seq: Replicated data

- A simple method for testing differential expression that incorporates within-group (or within treatment) variability relies on a **Generalized Linear Model (GLM)** with overdispersion -*var(Y)>E(Y)*-.
- It is a flexible generalization of ordinary **linear regression** that allows for response variables that have **other than a normal distribution**. (Normal distribution is one of the assumptions underlying linear regression)
- When data is counts of events (or items) then a **discrete distribution** (like Poisson) is more appropriate than approximating with a continuous distribution (Negative counts do not make sense).

RNA-seq: Replicated data

- A simple method for testing differential expression that incorporates within-group (or within treatment) variability relies on a **Generalized Linear Model (GLM)** with overdispersion -*var(Y)>E(Y)*-.
- It is a flexible generalization of ordinary **linear regression** that allows for response variables that have **other than a normal distribution**. (Normal distribution is one of the assumptions underlying linear regression)
- When data is counts of events (or items) then a **discrete distribution** (like Poisson) is more appropriate than approximating with a continuous distribution (Negative counts do not make sense).

RNA-seq: Balanced Block Designs

- Without careful planning an unblocked design faces a fundamental problem with generalizing the results: the potential for **confounding**.
- If the treatment effects are not separable from possible confounding factors, then for any given gene, there is no way of knowing whether the observed difference in abundance between treatment groups is due to biology or the technology.

RNA-seq: Balanced Block Designs

- Without careful planning an unblocked design faces a fundamental problem with generalizing the results: the potential for **confounding**.
- If the treatment effects are not separable from possible confounding factors, then for any given gene, there is no way of knowing whether the observed difference in abundance between treatment groups is due to biology or the technology.

RNA-seq: Balanced Block Designs

Example:

All replicates of treatment 1 are sequenced in lane 1 and all replicates of treatment 2 in lane 2, and goes on.

Any differences in expression between T1 and T2 are confounded with possible lane effects that may persist across flow cells.

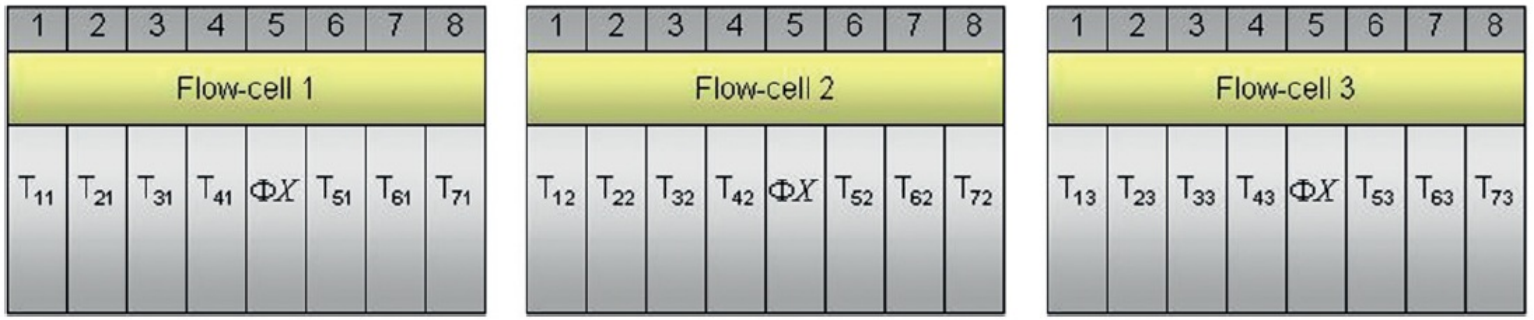


FIGURE 3.—A multiple flow-cell design based on three biological replicates within seven treatment groups. There are three flow cells with eight lanes per flow cell. The control ΦX sample is in lane 5 of each flow cell. T_{ij} refers to the jth replicate in the ith treatment group ($i = 1, \ldots, 7; j = 1, \ldots, 3$).

RNA-seq: Balanced Block Designs

- Different genes have different variances and are potentially subject to different errors and biases.
- Sources of variation affecting only a minority of genes should be integrated into the design as well (PCR-based GC bias). Complexity of the library.
- Two main sources of variation that may contribute to confounding effects:
  - **Batch effects:** errors that occur after random fragmentation of the RNA until it is imput to the flow cell (PCR, reverse transcription).
  - **Lane effects:** erros that occur from the flow cell until obtaining the data from the sequencing machine (bad sequencing cycles, base-calling)

RNA-seq: Balanced Block Designs

- Different genes have different variances and are potentially subject to different errors and biases.
- Sources of variation affecting only a minority of genes should be integrated into the design as well (PCR-based GC bias). Complexity of the library.
- Two main sources of variation that may contribute to confounding effects:
  - **Batch effects:** errors that occur after random fragmentation of the RNA until it is imput to the flow cell (PCR, reverse transcription).
  - **Lane effects:** erros that occur from the flow cell until obtaining the data from the sequencing machine (bad sequencing cycles, base-calling)

RNA-seq: Balanced Block Designs

- Different genes have different variances and are potentially subject to different errors and biases.
- Sources of variation affecting only a minority of genes should be integrated into the design as well (PCR-based GC bias). Complexity of the library.
- Two main sources of variation that may contribute to confounding effects:
  - **Batch effects:** errors that occur after random fragmentation of the RNA until it is imput to the flow cell (PCR, reverse transcription).
  - **Lane effects:** erros that occur from the flow cell until obtaining the data from the sequencing machine (bad sequencing cycles, base-calling)

RNA-seq: Barcoding

DNA fragments can be labeled or barcoded with **sample specific sequences** that allow multiple samples to be included in the same sequencing reaction while maintaining with high fidelity sample identities downstream.

Multiplexing can be used as a **control quality feature**, apart of increasing the number of samples per sequencing run, it offers the flexibility to construct balanced blocked designs for the purpose of testing differential expression.

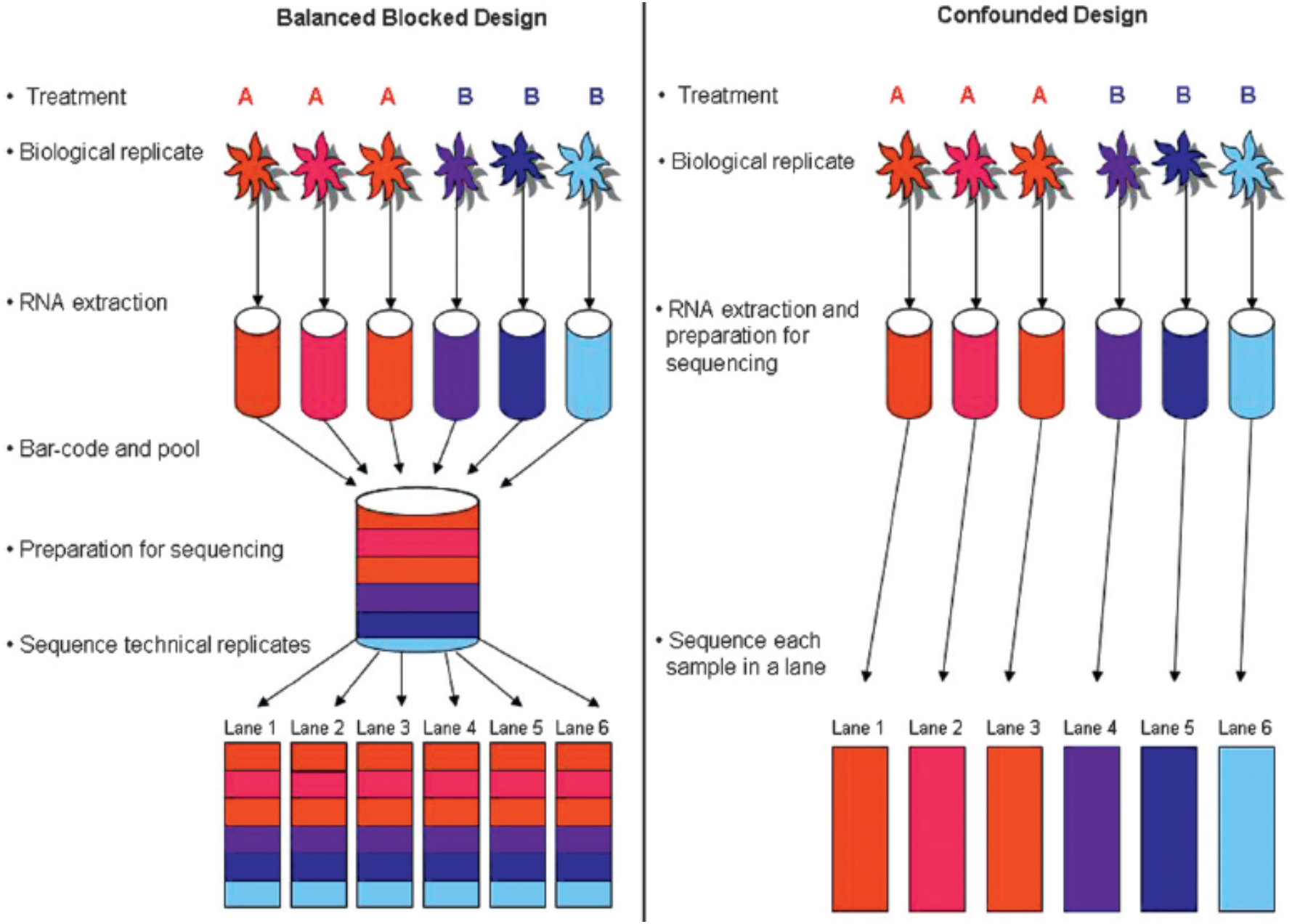
RNA-seq: Barcoding

DNA fragments can be labeled or barcoded with **sample specific sequences** that allow multiple samples to be included in the same sequencing reaction while maintaining with high fidelity sample identities downstream.

Multiplexing can be used as a **control quality feature**, apart of increasing the number of samples per sequencing run, it offers the flexibility to construct balanced blocked designs for the purpose of testing differential expression.

RNA-seq: Balanced Block Designs

- All the samples of RNA are pooled into the same batch and then sequenced in one lane of a flow cell.
- Any batch effects are the same for all the samples, and all effects due to lane will be the same for all samples.
- This can be achieved barcoding the RNA immediately after fragmentation


Balanced Blocked Design
• Treatment
A A A B B B
• Biological replicate
• RNA extraction
• Bar-code and pool
• Preparation for sequencing
• Sequence technical replicates
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6
Confounded Design
• Treatment
A A A B B B
• Biological replicate
• RNA extraction and preparation for sequencing
• Sequence each sample in a lane
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6

Balanced incomplete block designs (BIBD) and blocking without multiplexing

- In reality, technical constraints and the scientific hypotheses under investigation will dictate:
  - the number of treatments (**I**),
  - the number of biological replicates per treatment (**J**),
  - the number of unique barcoded (**s**) that can be included in one lane,
  - The number of lanes available for sequencing (**L**)

  When s<I a complete block design is not possible

If T is the total number of possible technical replicates , then a BIBD satisfies T=sL/JI.

Illumina has at the moment 12 different barcodes in a single lane, then in total 96 samples can be multiplexed.

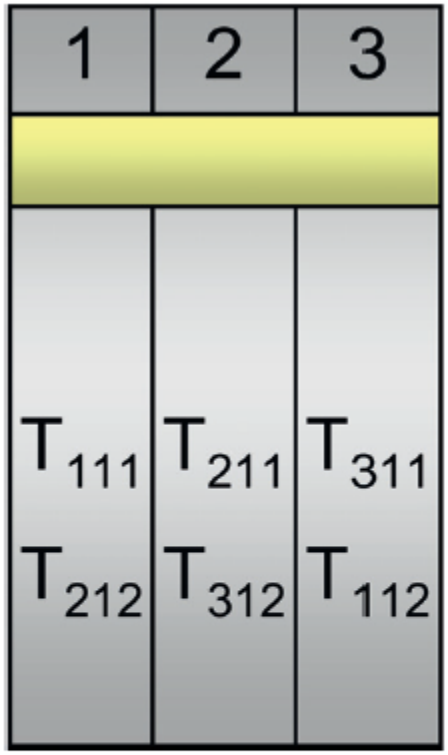


FIGURE 5.—A balanced incomplete block design (BIBD) for three treatment groups (T_1, T_2, T_3) with one subject per treatment group (T_{11}, T_{21}, T_{31}) and two technical replicates of each (T_{111}, T_{112}, T_{211}, T_{212}, T_{311}, T_{312}). After fragmentation, each of the three samples is bar coded and divided in two (*e.g.*, T_{11} would be split into T_{111} and T_{112}) and then pooled and sequenced as illustrated (*e.g.*, T_{111} is pooled with T_{212} as input to lane 1).

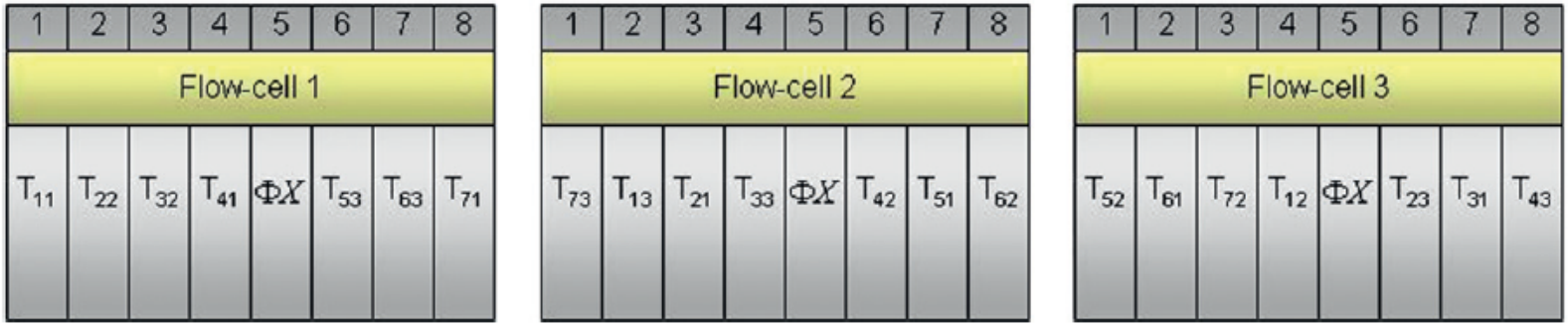


FIGURE 6.—A design based on three biological replicates within seven treatment groups. For each of the three flow cells there are eight lanes per flow cell and a control (ΦX) sample in lane 5. T_{ij} refers to the jth replicate in the ith treatment group ($i = 1, \ldots, 7; j = 1, \ldots, 3$). In this design the flow cells form balanced complete blocks, and the lanes form balanced incomplete blocks.

Simulations

2 treatments with T_{ijk}, where

i= treatment,

j= biological replicate and

k= technical replicate

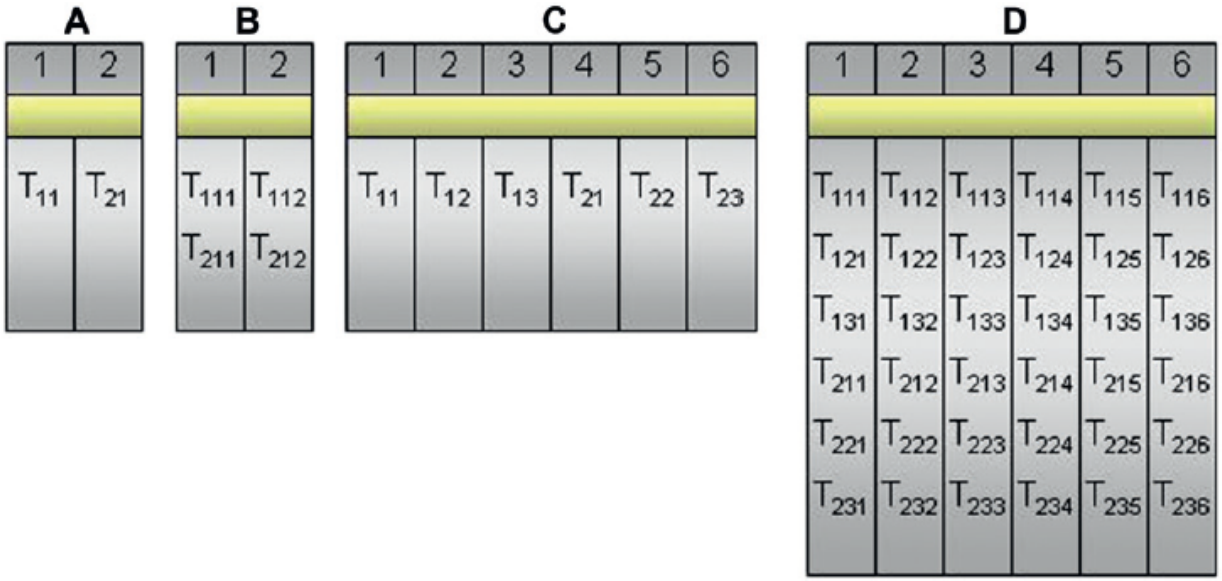
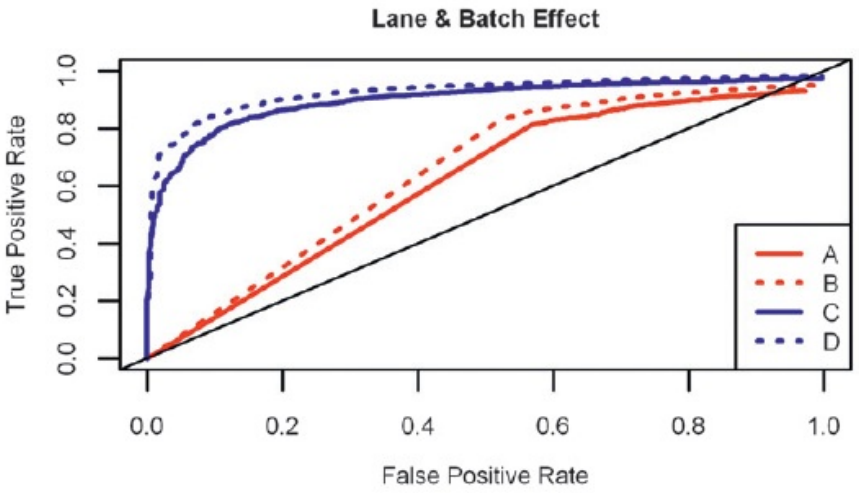
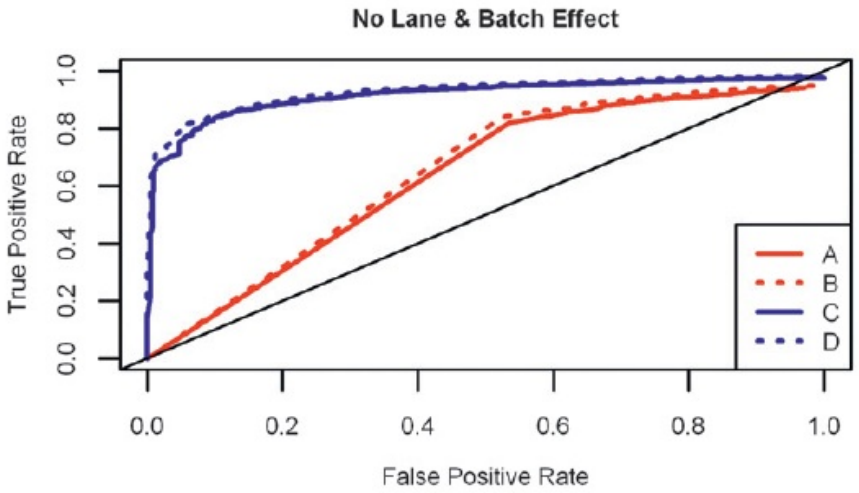
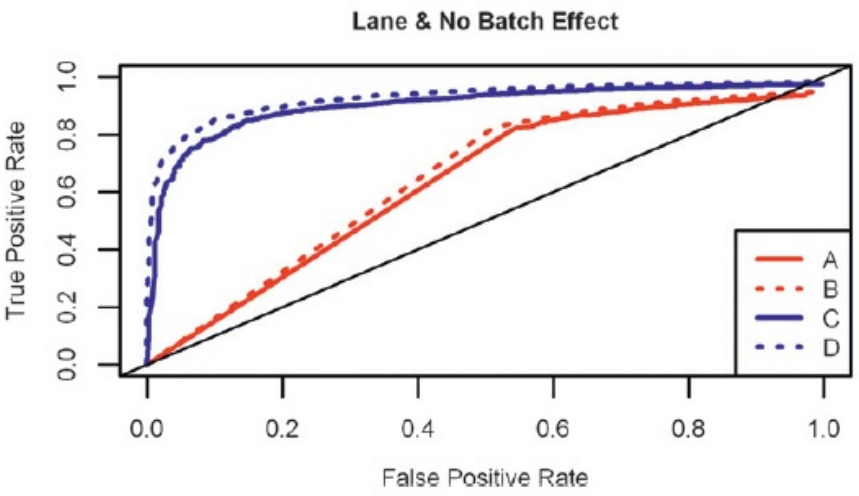


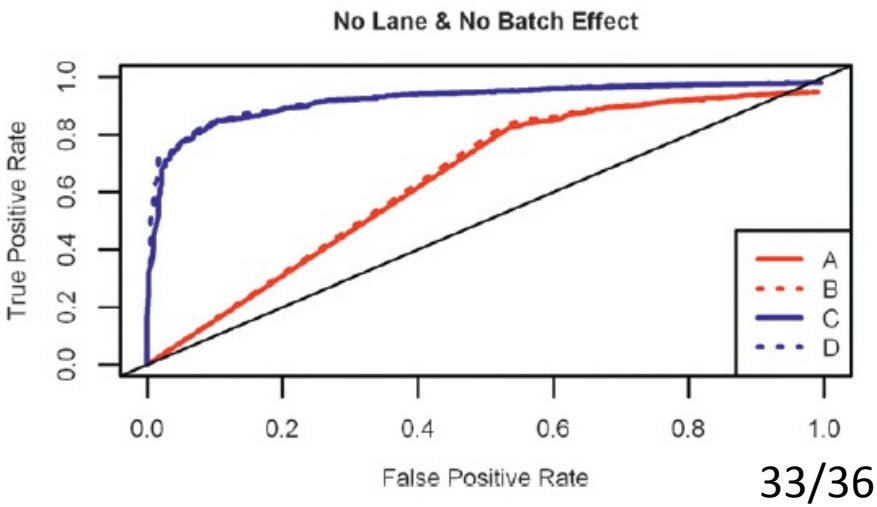
FIGURE 7.—Four designs (A–D) are compared in the simulation study for treatments T_1 and T_2. Design A is a biologically unreplicated unblocked design with one subject for treatment group T_1 (T_{11}) and one subject for treatment group T_2 (T_{21}). Design B is a biologically unreplicated balanced block design with T_{11} split (bar coded) into two technical replicates (T_{111}, T_{112}) and T_{21} split into two technical replicates (T_{211}, T_{212}) and input to lanes 1 and 2. Design C is a biologically replicated unblocked design with three subjects from treatment group T_1 (T_{11}, T_{12}, T_{13}) and three subjects from treatment group T_2 (T_{21}, T_{22}, T_{23}). Design D is a biologically replicated balanced block design with each subject (*e.g.*, T_{11}) split (bar coded) into six technical replicates (*e.g.*, T_{111}, . . . , T_{116}) and input to six lanes.

- Gene counts were simulated across treatment groups
- Compared the false positive rate (type I error, specificity) and the true positive rate (sensitivity/statistical power).









Discussion

- Replication, randomization and blocking are essential components of any well planned and properly analyzed design.
- NGS platforms allow us to work with the concepts of randomization and blocking (multiplexing).
- Biological replicates remains in the decision of the scientist.

Discussion

- The best way to ensure reproducibility and accuracy of results is to **include independent biological replicates** (technical replicates are not substitute) and to acknowledge anticipated nuisance factors in the design.
- **Balanced Block Designs** are as good as, if not better than, their unblocked counterparts in term of power and type I error and are considerable better when **batch and/or lane effects** are present.

References

- P. L. Auer and R. W. Doerge. 2010. *Statistical design and analysis of RNA sequencing data*. Genetics 185:405-416.

- R. A. Fisher. 1935. *The design of experiments*. Ed.2. Oliver & Boyd, Edinburgh.

- Photo of the cover obtained in: http://fc08.deviantart.net/fs70/f/2010/070/e/e/Colour_Bars___Melt_by_NehpetsDnalb.jpg