Design of Experiment – Part II

High-throughput Sequencing Workshop

Department of Biostatistics & Bioinformatics
Duke University Medical Center

July 17, 2017

Outline

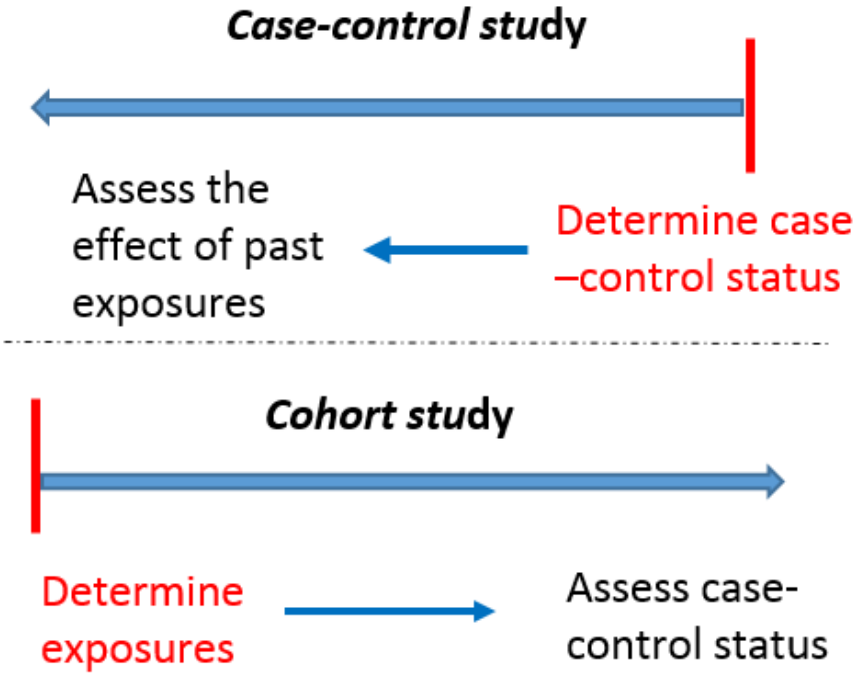
- Study design for clinical research
- Study design for genetic association studies.
- Study design related to gene expression studies – RNA-Seq
  - Technical variability in RNA-Seq
  - Experimental Designs in RNA-Seq

Main References:

1. Marioni et al. Genome Research 2008, 18:1509-17
2. McIntyre et al. BMC Genomics 2011, 12:293
3. Auer et al. Genetics 2010, 185:405-416

Study designs in clinical research

Epidemiology and observational study



- **case-control study:** Cases and controls are identified first.
- **cohort study:** Prospective observational study – Identify exposures (treatments) of interest first. Then, follow up subjects over time to assess their effects.
- **cross-sectional study:** Focus on a single time point or time interval. Exposures and outcomes are determined at the same time.

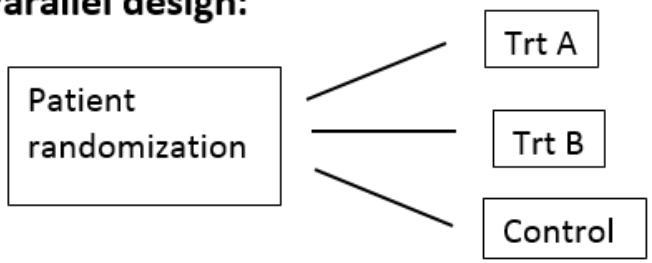
No randomization

Clinical trial

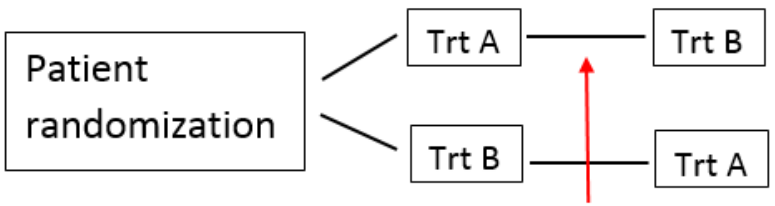
- The clinical trial design follows the EOD principles mostly same as the classical experimental design, except
  - It is harder to control various sources of variability comparing to laboratory, agriculture, and industrial experiments.
  - The variability of response variable may be larger than those from genetically identical animal or plants.
  - The enrollment of patients may be lengthy.
  - Subjects may drop out during the study.
- Need to satisfy ethical requirement
- Common used strategies: randomization, blocking, and blinding

Examples of clinical trial design

Parallel design:



Cross-over design



Wash out period

Factorial Design

Trt A and B	**Trt A only**
Trt B only	**Control**

e.g. investigating two drugs, A and B.

- Parallel design: Subjects are randomized to one of two or more arms. Each arm receive a different treatment.
- Cross-over design: Each patient gets both drugs and serve as his own control. Less sample size.
- Factorial design: For two or more intervention. Allow study interactive effect.

Genetics Association Studies

Genetics association studies

- **Objective:** Search genes/genetic variants that explain the susceptibility or causality of a disease or outcome.
- **Bio-Specimen:** DNA
- **Approaches:**
  - *Candidate gene approach:* Genotyping htSNPs in the genes or sequencing the entire genes of interest
  - *Genome wide association studies:* Genotyping high density SNPs across the genome.
  - *Whole exome sequencing or whole genome sequencing:* utilize high throughput sequencing (HTS) technology.

Biased and unbiased approaches

- Biased approach: Candidate gene analysis
  - Biologic or positional hypothesis:
    Example: focus on functional candidate genes within a linkage region
- Unbiased approach:
  - Genome wide association study (GWAS)
    - Uses of htSNPs to exploit LD across the genome, and get close to causal SNP
    - Identifies potentially novel genes and pathways in disease etiology
  - Whole exome sequencing or whole genome sequencing
    - Can be used for a full genome wide study (*i.e.* sequence all samples). However, the cost is high.
    - Can be used for variant or gene discovery for the phenotype of interest by sequencing a subset of samples and then follow-up the target variants by a large dataset.

Background of disease genetics association

- **Concept of allelic association:**
  - Alleles A and B at two loci are associated if the event that a gamete carries A is not independent of the event that the gamete carries allele B.

    AB ab AB ab ab AB AB ab ab ab

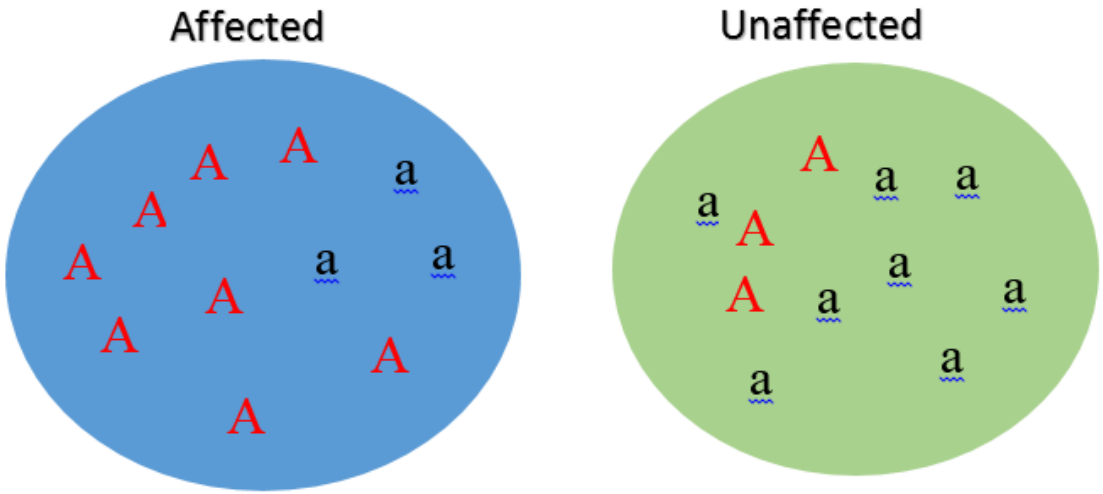
  - Alleles A and B are not associated if they occur in the gametes randomly.

    AB Ab AB aB ab aB Ab ab AB aB

  - Allelic association is population specific.
- **Linkage disequilibrium (LD)**: measure the degree of allelic association between two markers.

Background of disease genetic association

- **Disease/allele association**
  - Look for the association between markers and disease phenotype.
  - Allele **A** occurs more frequent in affected than unaffected subjects.

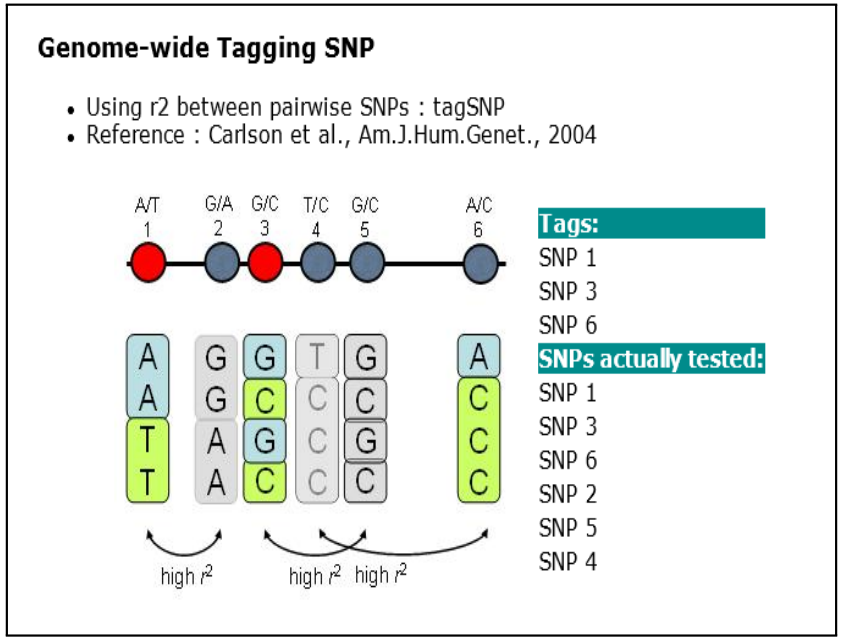


  - Assume the marker is in allelic association with the causal allele.

Background of disease genetics association

- LD decays with time and distance: $D_t = (1 - r)^t D_0$, where D_0 is the LD at the starting point, and r is the recombination rate.

- LD makes markers tightly linked, so we don't need to study every single SNP in the same LD block -use haplotype tagging SNP (htSNP) instead.
- Assume that 'significant' htSNP identified is in strong LD with the causal variant.



http://snp.istech21.com/snpanalyzer/2.0/overview/

Commonly used study designs

- Family-based approach
  - Recruit proband and his/her family members.
  - The unaffected family members or parents can serve as controls to evaluate variant association to the phenotype of interest.
  - **Types of family structure:** parents-offspring trio, discordant sibpair with or without parents, extended family
  - Apply to linkage studies: To identify chromosomal regions that harbor genes linked to the phenotype of interest.
  - Apply to association studies.
  - Often limited by the number of families one can recruit.
- case-control approach
  - Use unrelated population samples, easier to recruit, but selection of controls is important.
  - prone to spurious results by population admixture

Effect of Population admixture

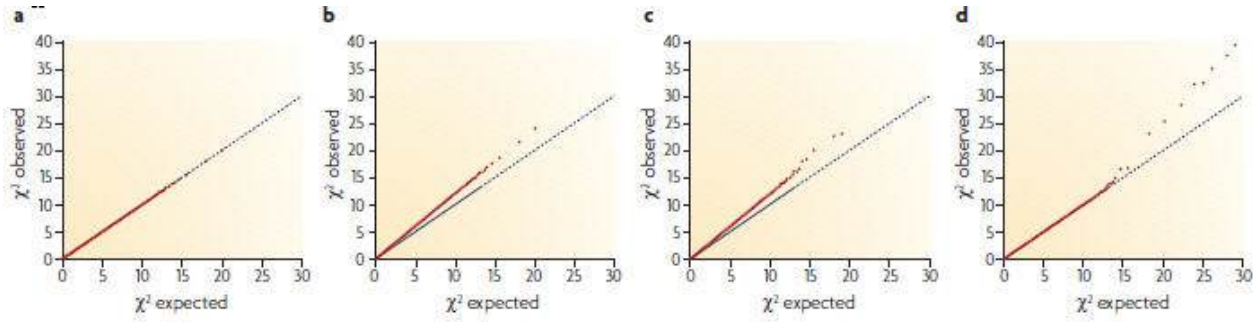
- Two (or more) mixing populations can lead to associations created due to the differences in allele frequencies in the mixing populations
  - **Population A:** **A** allele is VERY common; Disease allele (**D**) occurs randomly with **A** or **a** alleles.
  - **Population B:** **a** allele is VERY common; Disease allele (**D**) does not exist.
  - **Admixture population:** Assume equal mixed of populations A and B (allele **A** with frequency of 0.5), we will observed an association between A and D.
- Population structure will lead to false positive results for genetic association studies.

How to adjust population structure?

- **Family-based design**
- **Genomic control:** Estimate degree of population stratification by typing 20-60 unlinked markers on same cases and controls used for studying candidate gene association (Devlin and Roeder 1999)
- **Structure:** Alternative method based on explicit modeling of population structure (Pritchard and Rosenberg (1999))
- **Eigenstrat:** Use principal components (PCs) analysis to explicitly model ancestry differences between cases and controls. (Price et al. 2006)

The idea of using principal components to adjust for population structure can also be applied to gene expression data.

Visualization: Q-Q plots



Quantile-Quantile Plots compare observed distribution of test statistics to that expected under the null hypothesis of no association (McCarthy et al. 2008).

- No association, observed = expected
- Probably mostly population substructure, deviations across distribution
- Possible true associations, but also population substructure
- True association, deviations at the highest end of the distribution

Useful R code: http://www.broadinstitute.org/diabetes/scandinavs/figures.html

Gene expression studies

Technology for measuring gene expression

- **Objective:** Investigate differential expression of gene(s) due to different conditions (*e.g.* disease status, treatment effect)
- **Bio-specimen:** RNA
- **Methods for measuring expression levels:**
  - Reverse transcription polymerase chain reaction (RT-PCR): the most sensitive technique to detect and quantifying mRNA
  - Serial Analysis of Gene Expression (SAGE):
  - gene expression microarray: Probe-based chip.
  - RNA-Seq: sequencing of RNA molecules.
- SAGE, expression microarray, and RNA-Seq are all high throughput methods, which can detect genes across the genome.

RNA-Seq

Steps of a RNA-Seq experiment[1]

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

[1] **Auer et al. Genetics 2010.**

Sources of variability

Types of variability applying to any experiments

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

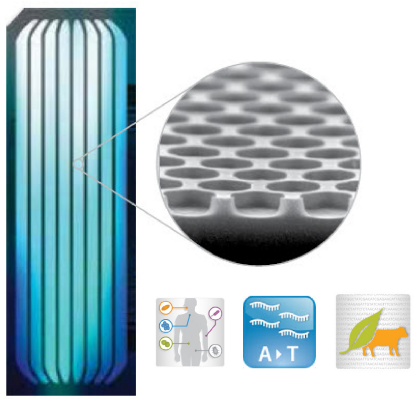
These sources of variability need to be considered in the experimental design.

Technical variability

Types of 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)



Designs of evaluating technical variability

Example I[2]

Methods

RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays

John C. Marioni,[1,6] Christopher E. Mason,[2,3,6] Shrikant M. Mane,[4] Matthew Stephens,[1,5,7] and Yoav Gilad[1,7]

[1]*Department of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA;* [2]*Program on Neurogenetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA;* [3]*Department of Genetics, Yale University School of Medicine, New Haven, Connecticut 06520, USA;* [4]*Keck Biotechnology Laboratory, New Haven, Connecticut 06511, USA;* [5]*Department of Statistics, University of Chicago, Chicago, Illinois 60637, USA*

[2]**Marioni et al. Genome Research, 2008.**

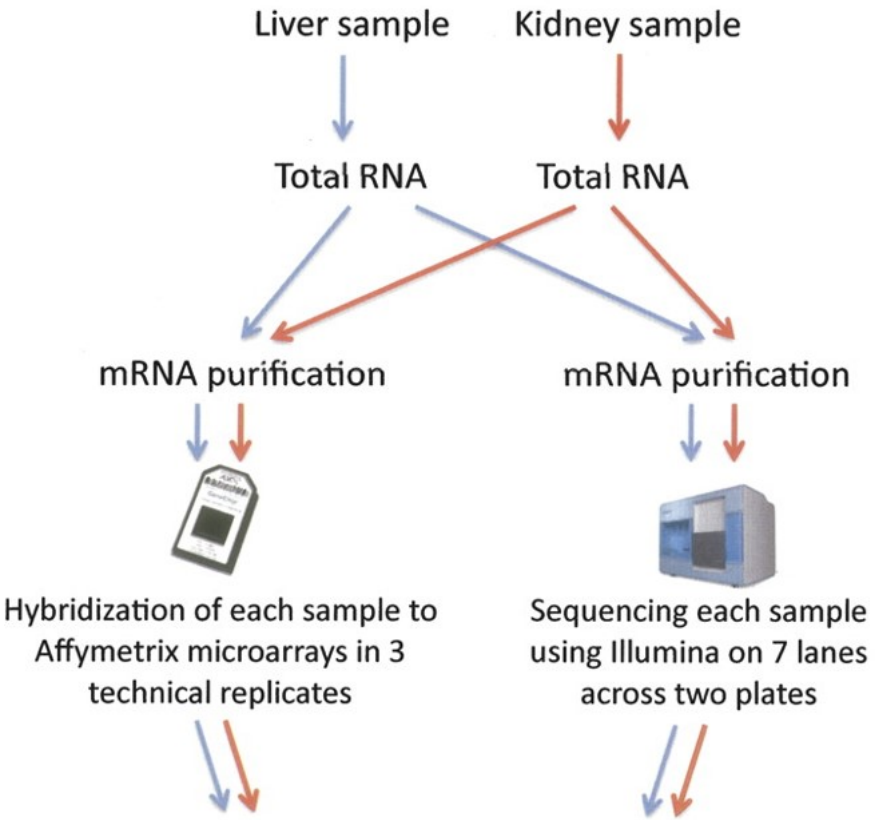
Designs of evaluating technical variability

Example I: Marioni et al.

- **Objective:** Assess the technical reproducibility of Illumina RNA-Seq
  - Comparison between platforms
  - Evaluate technical variability of RNA-Seq

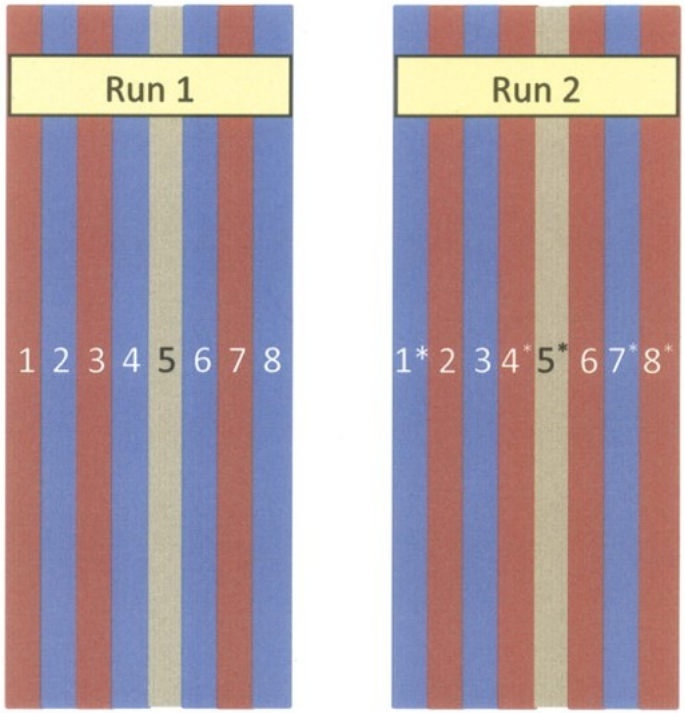
- **Outline of the experiment**
  - Two sequencing platforms: Illumina RNA-Seq (8 lanes) and Affymetrix microarray
  - Two samples: a liver and a kidney samples
  - Two cDNA concentration (3pM and 1.5pM)
  - One lane for a control sample.
  - Each sample were sequenced 7 times total in two flow-cell runs.

A



B

Illumina study design



Kidney
Liver

* Sequenced at a concentration of 1.5 pM

Design of evaluating technical variability

Example I: Marioni et al.:

- **Outline of the experiment**
  - Two sequencing platforms: Illumina RNA-Seq (8 lanes) and Affymetrix microarray
  - Two samples: a liver and a kidney samples
  - Two cDNA concentration (3pM and 1.5pM)
  - One lane for a control sample.
  - Each sample were sequenced 7 times total in two flow-cell runs.
- **What can they compare with this design?**

Design of evaluating technical variability

Example I: Marioni et al.:

- **Outline of the experiment**
  - Two sequencing platforms: Illumina RNA-Seq (8 lanes) and Affymetrix microarray
  - Two samples: a liver and a kidney samples
  - Two cDNA concentration (3pM and 1.5pM)
  - One lane for a control sample.
  - Each sample were sequenced 7 times total in two flow-cell runs.
- **What can they compare with this design?**
  - **Platform differences:** Two methods for gene expression

Design of evaluating technical variability

Example I: Marioni et al.:

- **Outline of the experiment**
  - Two sequencing platforms: Illumina RNA-Seq (8 lanes) and Affymetrix microarray
  - Two samples: a liver and a kidney samples
  - Two cDNA concentration (3pM and 1.5pM)
  - One lane for a control sample.
  - Each sample were sequenced 7 times total in two flow-cell runs.
- **What can they compare with this design?**
  - **Platform differences:** Two methods for gene expression
  - **Technical variability:** Same sample sequenced in different lanes, two separate runs

Design of evaluating technical variability

Example I: Marioni et al.:

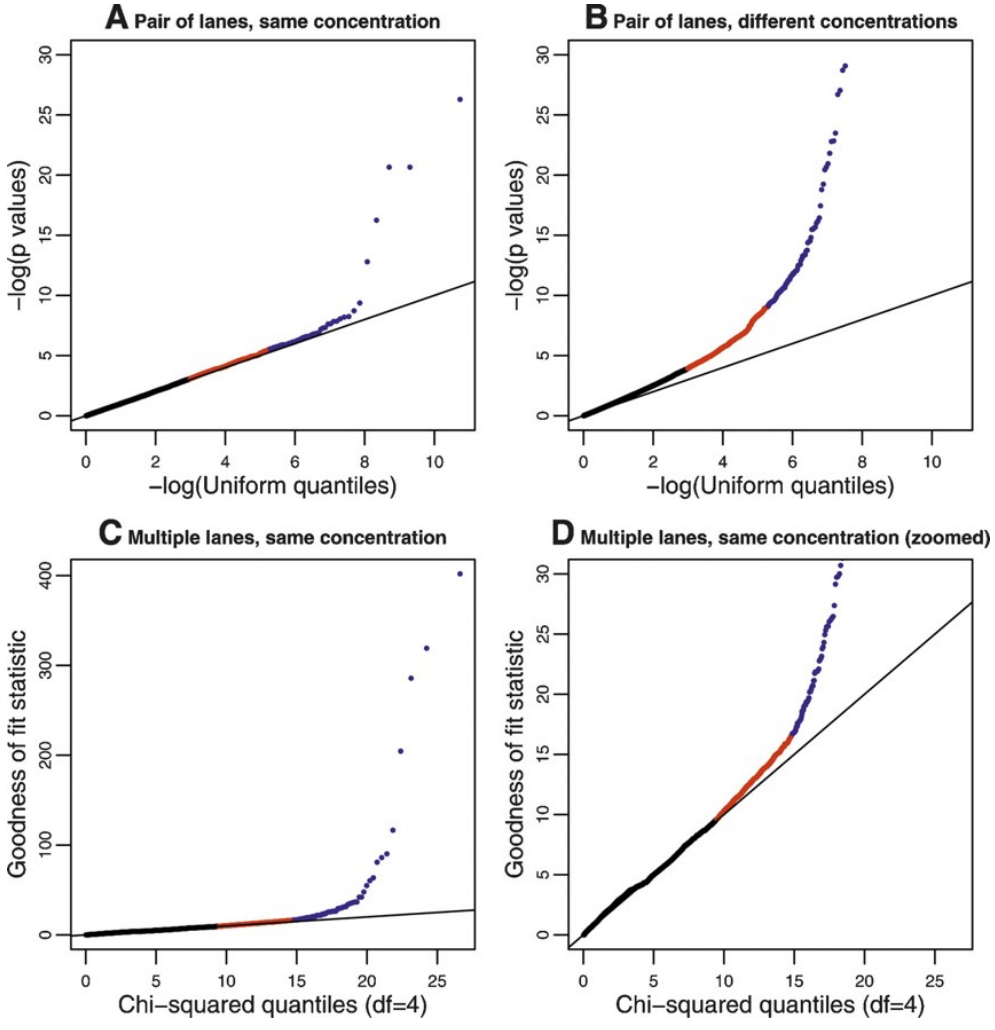
- **Outline of the experiment**
  - Two sequencing platforms: Illumina RNA-Seq (8 lanes) and Affymetrix microarray
  - Two samples: a liver and a kidney samples
  - Two cDNA concentration (3pM and 1.5pM)
  - One lane for a control sample.
  - Each sample were sequenced 7 times total in two flow-cell runs.
- **What can they compare with this design?**
  - **Platform differences:** Two methods for gene expression
  - **Technical variability:** Same sample sequenced in different lanes, two separate runs
  - **Effect of cDNA concentration:** two concentrations (3 and 1.5pM)

Design of evaluating technical variability

Example I: Marioni et al.:

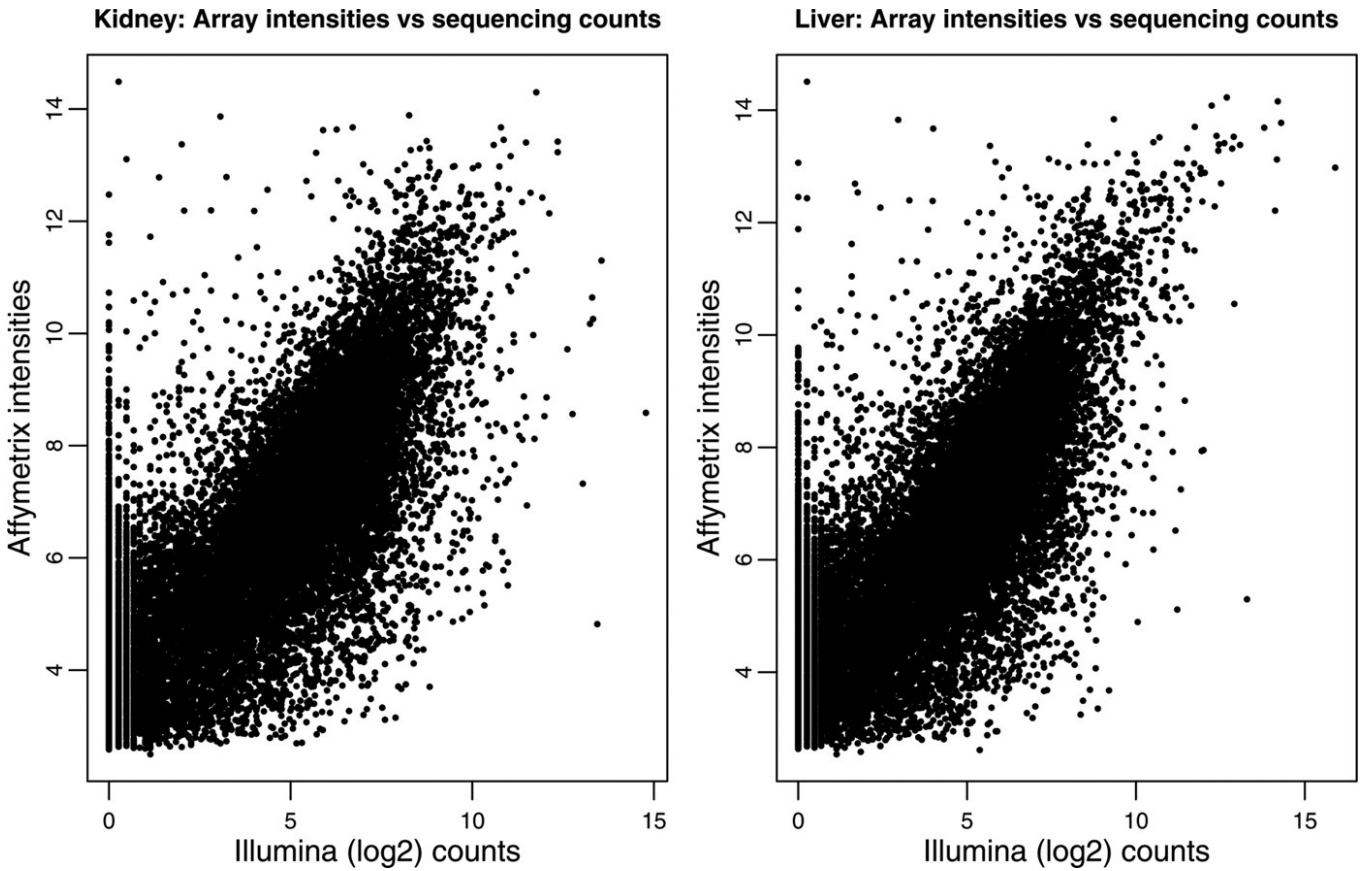
- **Outline of the experiment**
  - Two sequencing platforms: Illumina RNA-Seq (8 lanes) and Affymetrix microarray
  - Two samples: a liver and a kidney samples
  - Two cDNA concentration (3pM and 1.5pM)
  - One lane for a control sample.
  - Each sample were sequenced 7 times total in two flow-cell runs.
- **What can they compare with this design?**
  - **Platform differences:** Two methods for gene expression
  - **Technical variability:** Same sample sequenced in different lanes, two separate runs
  - **Effect of cDNA concentration:** two concentrations (3 and 1.5pM)
  - **Differential expression between liver and kidney tissues**: Two tissue samples

Plots for assessing lane effect



A: Same sample, same concentration; **B** Same sample, different concentration; **C&D:**Goodness-of-fit for Poisson distribution – kidney samples

Comparison between platforms Comparing counts from Illumina sequencing with normalized intensities from the array, for kidney (left) and liver (right).



Spearman correlation = 0.73 for liver, 0.75 for kidney

Conclusion and issues in the design

- Summary of Marioni et al. 2008
  - Illumina RNA-Seq is replicable and has advantage over microarray
  - **Lane effect is small. (??)**
  - Larger difference between runs **(Batch effect)**
  - larger difference between cDNA concentration
  - Suggested that it is OK to run one sample per lane
- **Issues in the Design: No replicates, one sample only**
  - Is it sufficient to use one sample per tissue type to conclude low lane effect?
  - Can we partition biological variation (*e.g.* liver vs. kidney) from technical variation?

Example II: McIntyre et al.[3]

McIntyre *et al. BMC Genomics* 2011, **12**:293
http://www.biomedcentral.com/1471-2164/12/293

RESEARCH ARTICLE Open Access

RNA-seq: technical variability and sampling

Lauren M McIntyre[1*], Kenneth K Lopiano[2], Alison M Morse[1], Victor Amin[1], Ann L Oberg[3], Linda J Young[2] and Sergey V Nuzhdin[4]

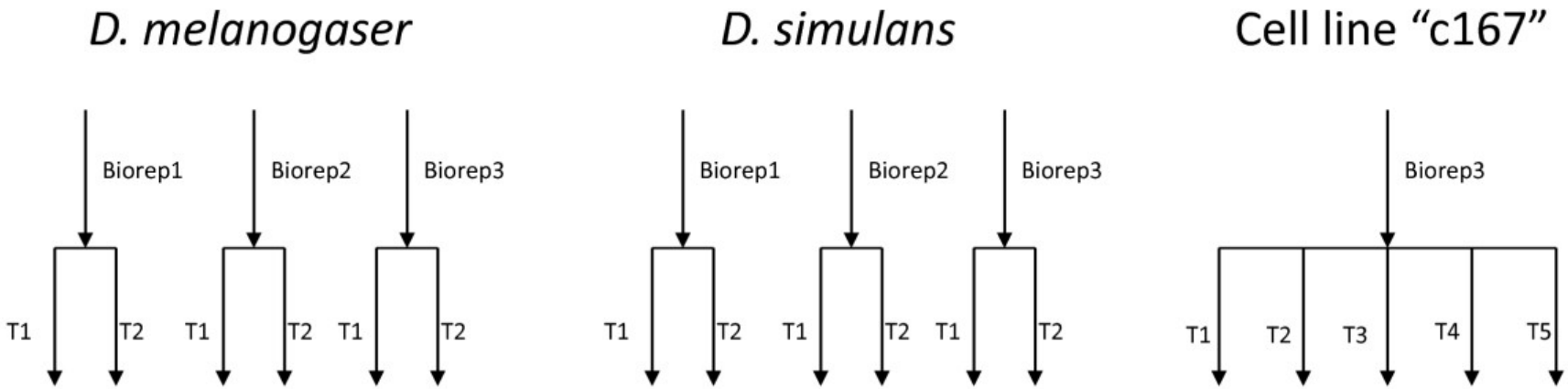
[3] **McIntyre et al. BMC Genomics 2011.**

Example II: McIntyre et al.
RNA-seq: technical variability and sampling

- **Objective:**
  - Does technical variability exist?
  - Is the impact of technical variability the same for all levels of coverage?

- **Experiments:**
  1. Three independent samples (*D. melanogaster* female), two technical replicates per sample, run on two lanes of a Solex/Illumina flow cell.
  2. Three independent samples (*D. simulans* male), two technical replicates per sample, run on two lanes of a flow cell.
  3. One sample (*D. melanogaster* cell lines), 5 replicates, run on 5 lanes of a flow cell.

All are 36 base-paired end. The relationship of lanes for cell lines (same or independent flow cells) is unknown.



D. melanogaster and *D. simulans* are single library run on multiple lanes. *D. melanogaster* c167 cell lines are not exactly the same library run.

Variation of sequence reads

Table 1 from McIntyre et al. 2011

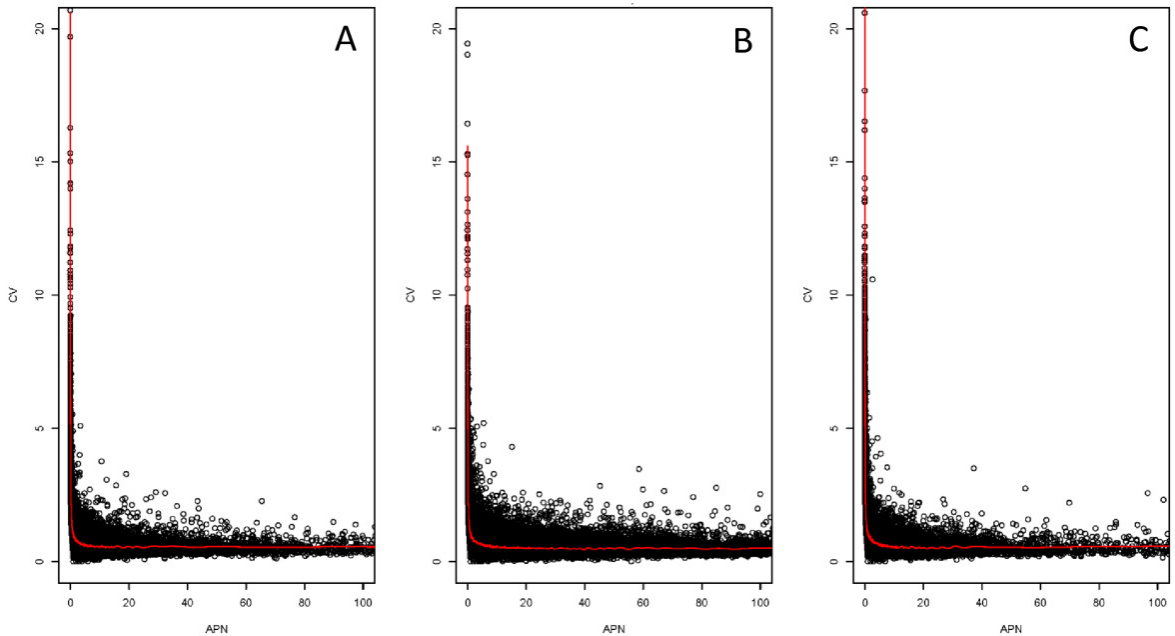
Table 1 Mappable reads per lane in each of the three experiments

Experiment	BR	TR	Mappable Reads	Exons detected	Exons with an average coverage of more than 5 reads per nucleotide	Contigs present in all samples of each experiment
c167	1	1	5888686	39156	13432	19248
c167	1	2	5951769	39202	13517	19248
c167	1	3	7146461	39954	15684	19248
c167	1	4	7544117	40201	16355	19248
c167	1	5	7377032	40120	16089	19248
D. sim.	1	1	5174398	45878	14517	20339
D. sim.	1	2	4979485	45808	13912	20339
D. sim.	2	1	27595266	51701	35303	20339
D. sim.	2	2	28691914	51857	35942	20339
D. sim.	3	1	27601233	51834	34968	20339
D. sim.	3	2	27748704	51822	35008	20339
D. mel.	2	1	10584341	48114	13396	17864
D. mel.	2	2	13399722	49073	19916	17864
D. mel.	3	1	12065885	48281	14794	17864
D. mel.	3	2	11794255	48319	17961	17864
D. mel.	4	1	10375138	47812	15718	17864
D. mel.	4	2	9283979	47460	14344	17864

Data variation seen between technical replicates and between biological replicates.

Visualization of the data

Coefficient of variation (CV) vs. average depth per nucleotide (APN); **APN:** within each lane, average number of reads per exon

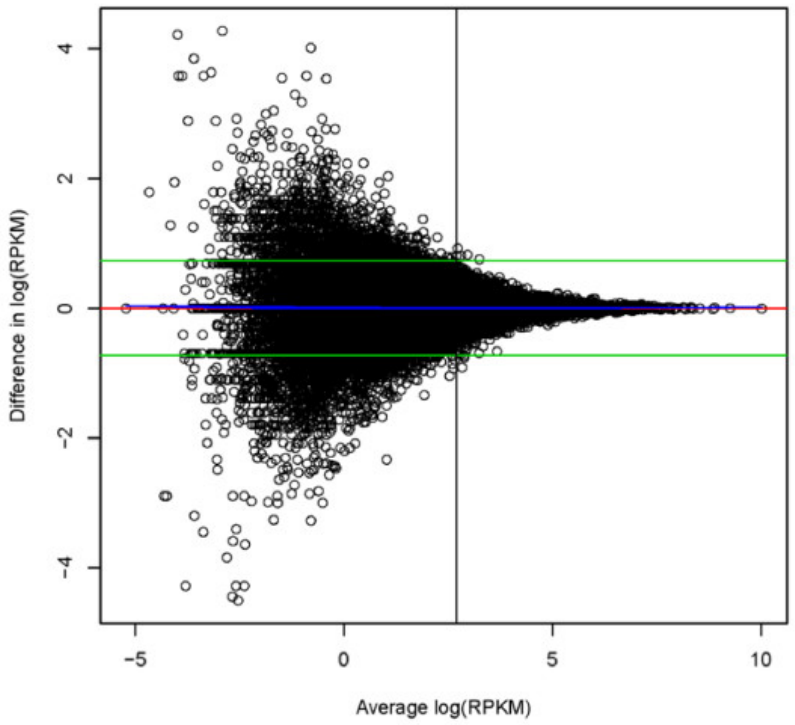


Lower coverage has higher variation

A: *D. simulans* BR2,TR2; **B:** *D. melanogaster* female heads BR2,, TR1; **C:** TR1 for cell line c167.

Visualization of the data

MA plot: Minus vs. average, Bland and Altman plot



Green line: One standard deviation region

Low expression level has higher disagreement

Summary of McIntyre et al. 2011

- **Technical variation exits:**
  - Mappable reads per lane various among the technical replicates
  - Inconsistent detection of exons between technical replicates: The number of exons detected increases with the number of mappable reads.
  - Agreement between technical replicates varies: kappa ranges from 0.63-0.81
  - Higher variability for those with low coverage (< 5 reads per nucleotide) or low expression level
  - Random sampling of total RNA reads (*e.g.* 0.0013% of 30 millions reads) may contribute to the variability.
- **Biological variation is larger than technical variation.**
- **Suggestion:** Inclusion of technical replicates is as important as biological replicates. Multiplexing design can eliminate the lane effect for a small experiment.

Designs for RNA-Seq

Designs for RNA-Seq

Reference paper:[4]

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

[4] **Auer and Doerge, Genetics, 2010.**

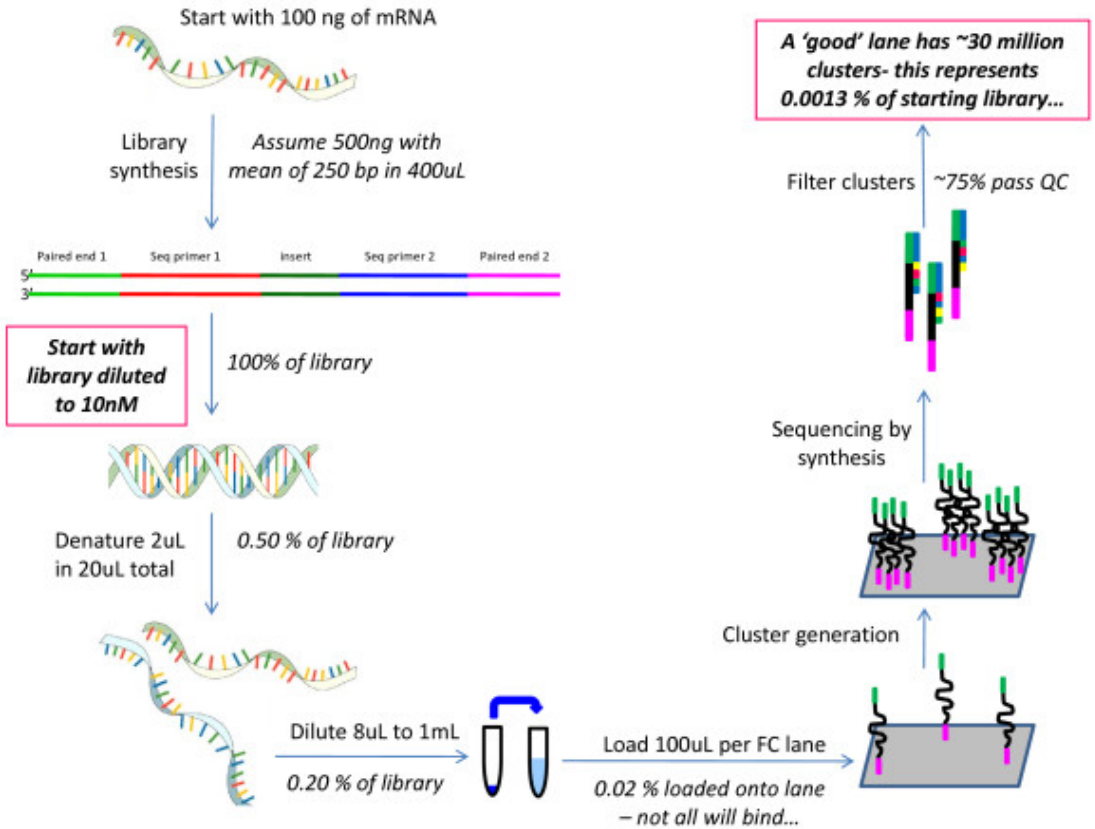
Unreplicated data

- **Three levels of sampling in RNA-Seq:**
  - Subject sampling
  - RNA sampling
  - Fragments sampling
- **Unreplicated data**:
  - Mostly from observational studies
  - In the case of RNA-Seq
    - No biological replicates
    - One sample per treatment group
- **Problem:** This design can investigate only the differences derived from RNA and fragment-level sampling, but not subject sampling.

More on 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 fragements loaded into a flow cell resulted in fragment sampling.

More on RNA and fragment sampling

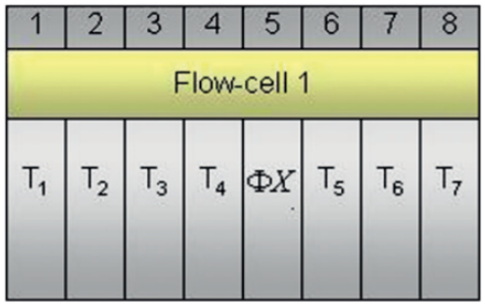


Library concentration 10nM=4pM $\rightarrow$ $\frac{4}{10^{12}} \times 6.02 \times 10^{23} = 2.408 \times 10^{12}$ total molecules in the library $\rightarrow$ $\frac{30,000,000}{2.408 \times 10^{12}} = 0.0013\%$ of molecules to be analyzed.

Unreplicated data

Outline of experiment:

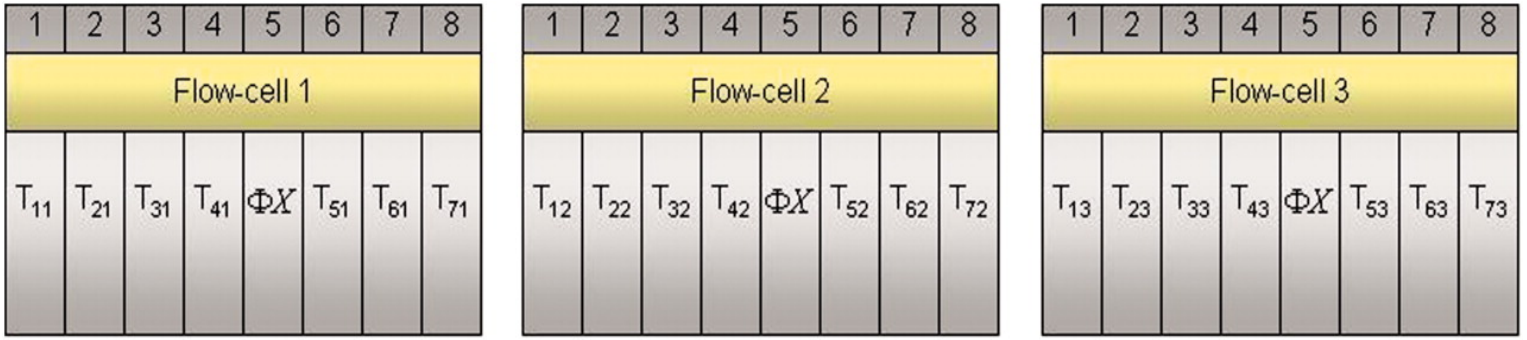
- mRNA isolated from subjects within different treatment group $(T_1, \cdots, T_7)$.
- a ΦX genomic sample is loaded to lane 5 as a control
- ΦX can be used to recalibrate the quality score of sequencing reads from other lane.



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



- **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.

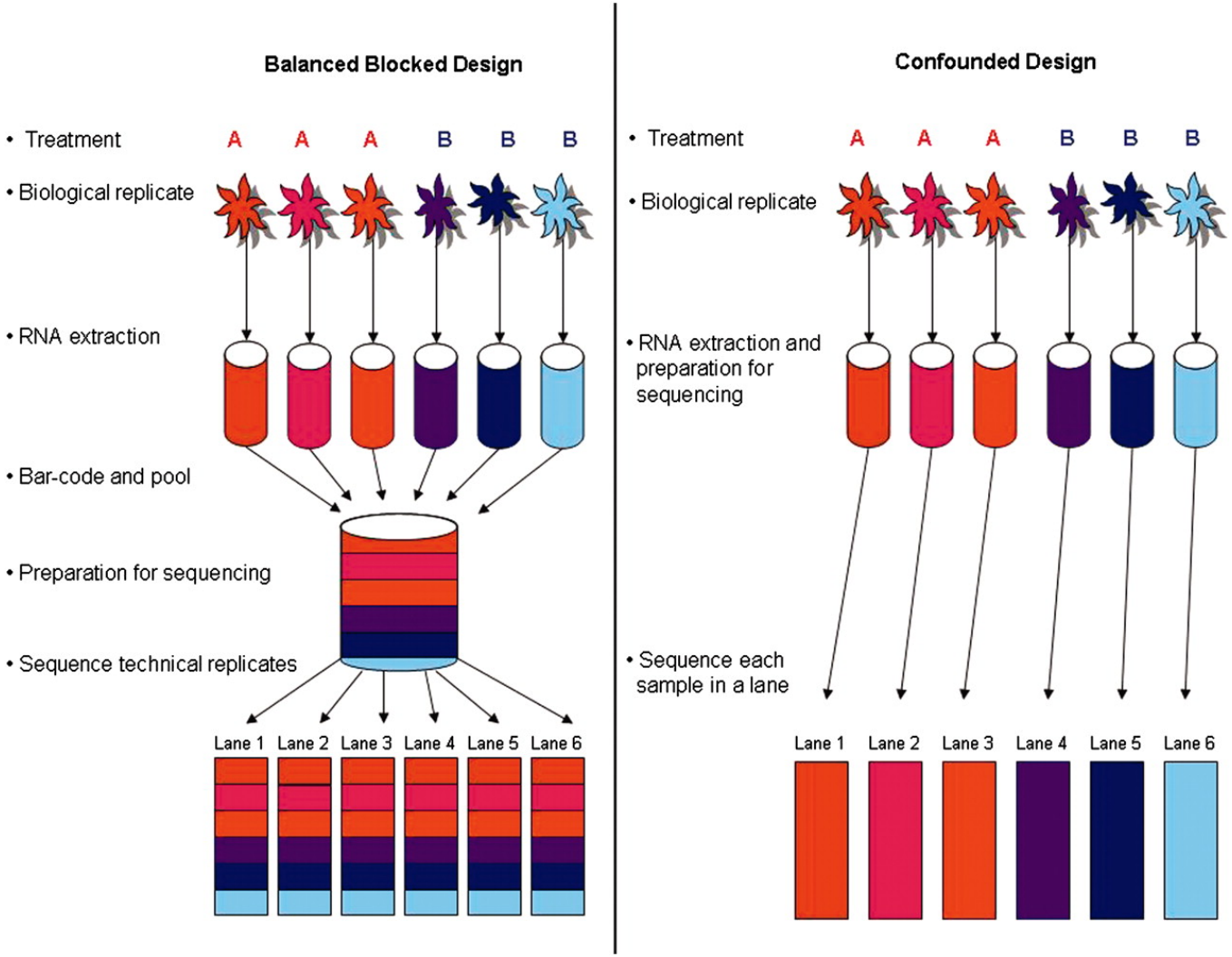
Balanced block design

- **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

Balanced Block Design - I

- Three biological replicates per treatment ($j = 1, \cdots, 3$)
- treatment group (A and B) ($i = 1, \cdots, 2$)
- 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

Balanced Block design - I



Analysis model for BBD I

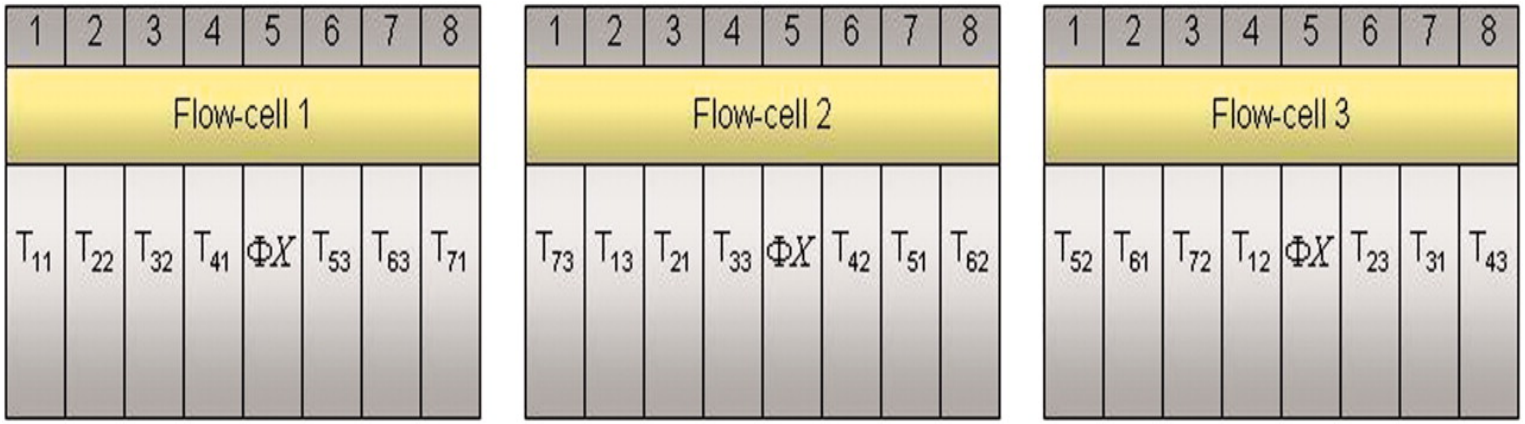
- **Dependent variable:** DGE measures, defined by the distribution you assumed for the sequence reads. For example,
  - Auer et al. assumed $y_{ijk} \sim Possion(\mu_{ijk})$.
  - DESeq2 uses Negative Binomial model.

  In Auer et al, $y_{ijk} = \sum_t y_{ijkt}$, 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}).

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

- 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.

Balanced block design II-without multiplexing



- 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 design 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.

Balanced incomplete block design (BIBD)

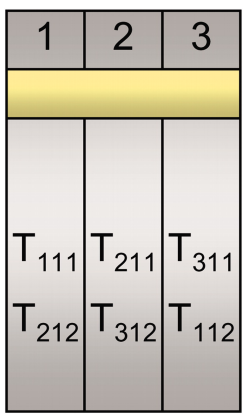
Assume:

- number of treatment (I)
- number of biological replicates per treatment (J)
- number of unique barcodes (s) that can be included in one lane
- number of lanes available for sequencing (L)

If the number of unique bar codes (s) in one lane is less than the number of treatments ($s < I$), balanced block design is impossible.

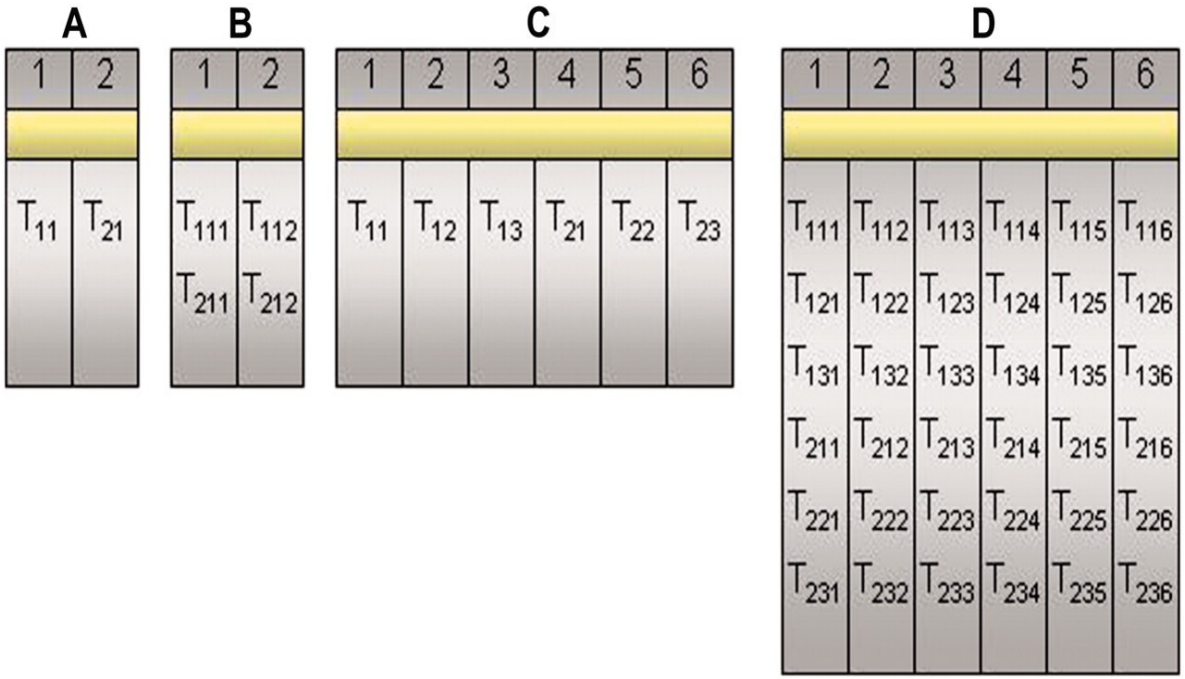
BIBD

- For a given number of treatment groups (I), sample per treatment group (J), unique barcodes (s), and number of available lanes (L), the total number of technical replicates (T) in BBID is $T = \frac{sL}{IJ}$.
- **Example of BIBD:**
  - Assume 3 treatment group ($i = 3$), one subject per treatment group ($j = 1$), two unique barcodes ($s = 2$), and three available lanes ($L = 3$).
  - The total number of technical replicates is $T = \frac{2 \times 3}{3 \times 1} = 2$.

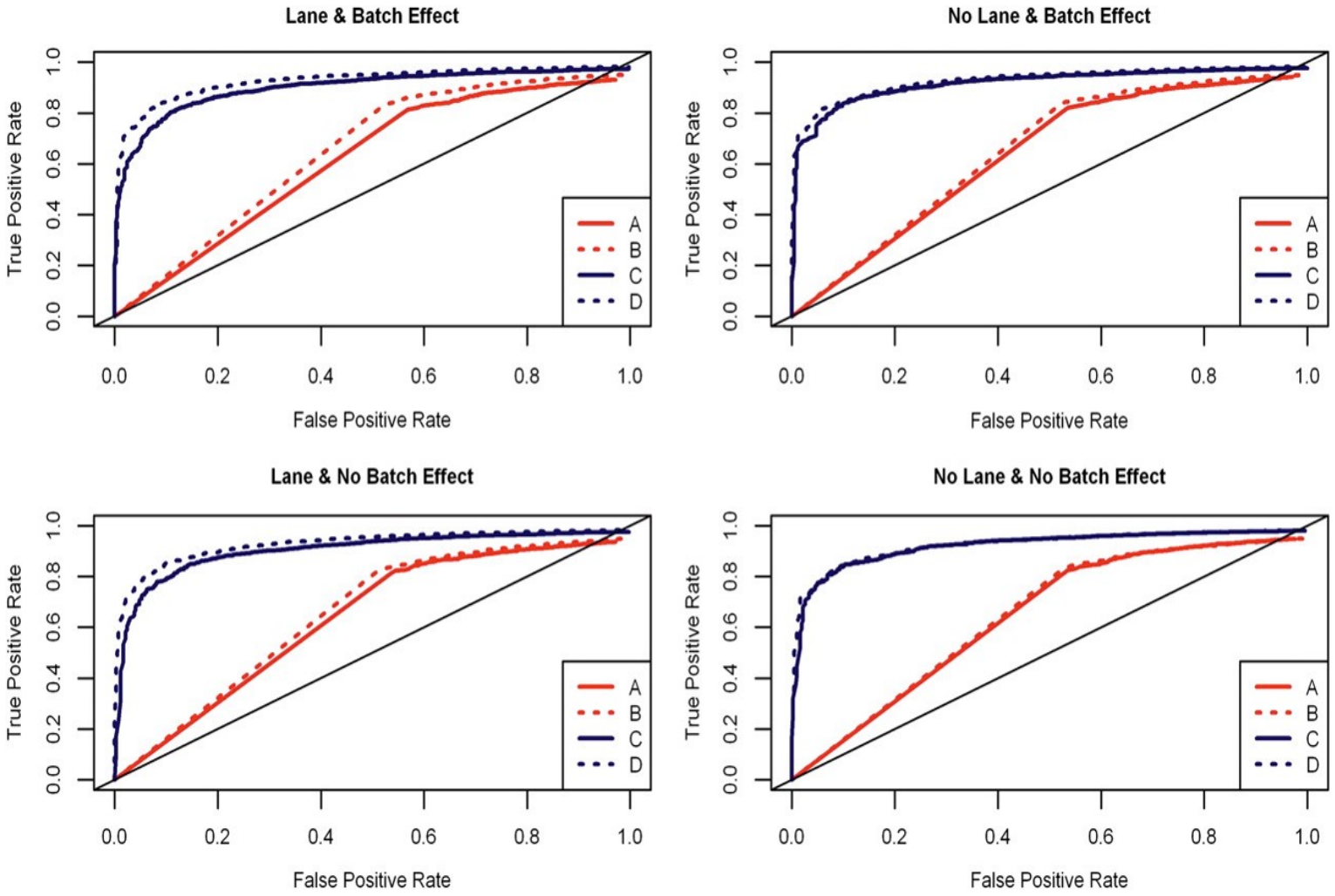


- T_{ijk} is for treatment i, subject j, and technical replicates t.
- For Illumina, a total of 12 unique barcodes can be used in one lane. Therefore, 96 samples can be multiplexed in one flow-cell run.

Performance comparison between designs by simulation studies



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

- The classical principals of experimental design still apply to RNA-Seq
- Technical variation exists and should be taken into account.
  - Lane effect, batch effect
- RNA-Seq data consist of variation from subject sampling, RNA sampling, RNA fragment sampling
- Multiplexing in NGS allow us to implement randomization and blocking.
- Take advantages of visualization tools (*e.g.* scatter plots, MA plots, QQ plots) to learn your data.
- When you deal with human data for genetic study, make sure examining the effect of population structure.