Designing Functional Genomics Experiments for Successful Analysis

RORY STARK
18 SEPTEMBER 2017

Agenda

WHY PERFORM EXPERIMENTS?

WHY THINK ABOUT EXPERIMENTAL DESIGN?

WHAT MAKES FOR A WELL DESIGNED EXPERIMENT?

KEY ASPECTS OF EXPERIMENTAL DESIGN

- Experimental variables

- Power: variance and replicates

- Bias: confounding factors, randomisation, and controls

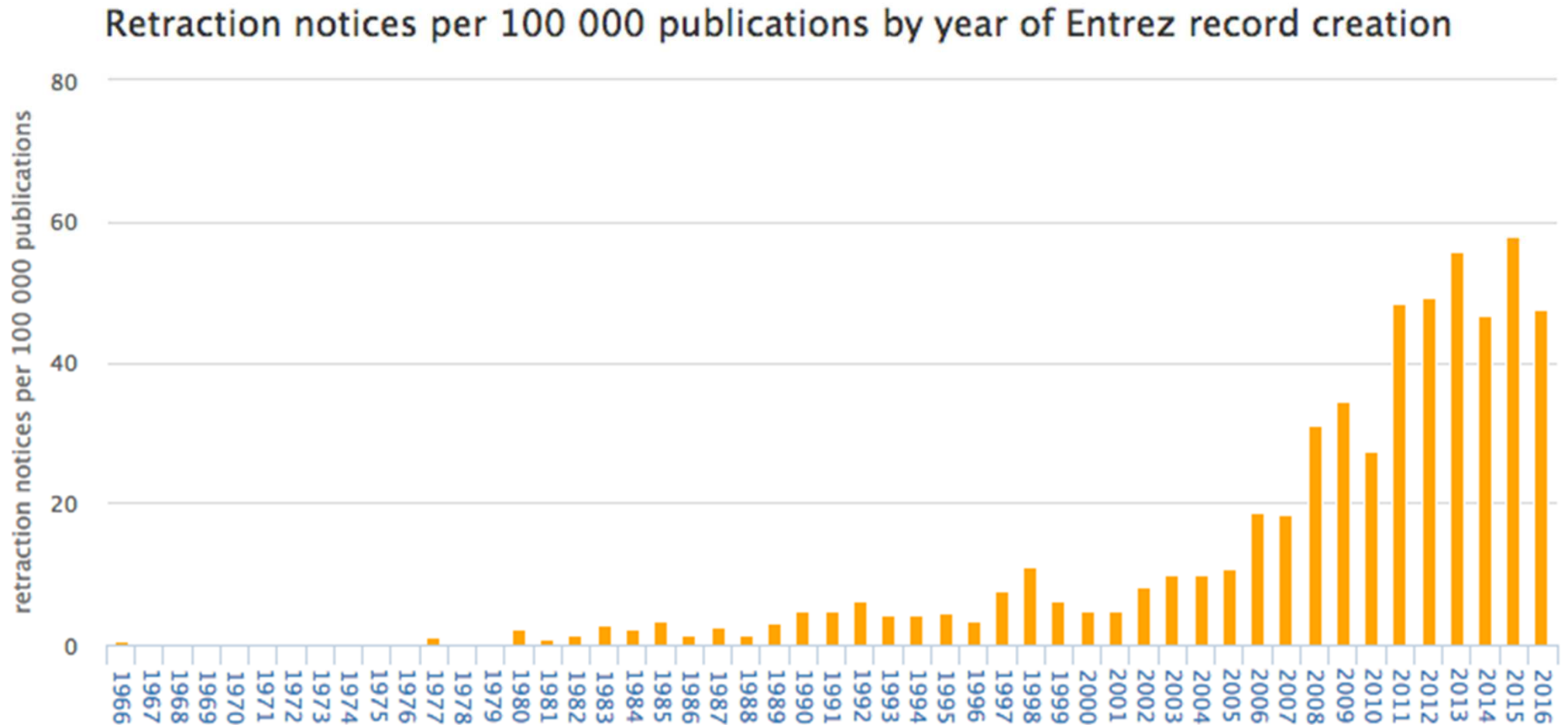**DESIGN PARAMETERS FOR FUNCTIONAL SEQUENCING EXPERIMENTS**

EXPERIMENTAL DESIGN PROCESS AT CRUK-CI

Why Perform Experiments?

Why Think About Experimental Design?

Crisis in Reproducible Research

Retraction notices per 100 000 publications by year of Entrez record creation

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

47 of 53 high-profile cancer studies were not reproducible!

Drug development: Raise standards for preclinical cancer research

C. Glenn Begley & Lee M. Ellis

Affiliations | Corresponding author

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Need for Good Design

Consequences of Poor Experimental Design…

- **Cost** of experimentation. We have a responsibility to CRUK donors!

- **Limited & Precious** material, esp. clinical samples.

- **Immortalization** of data sets in public databases and methods in the literature. Our bad science begets more bad science.

- **Ethical concerns** of experimentation: animals and clinical samples.

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

A Well-Designed Experiment:

Should have

CLEAR OBJECTIVES

FOCUS AND SIMPLICITY

SUFFICIENT POWER

RANDOMISED COMPARISONS

And be

PRECISE

UNBIASED

AMENABLE TO STATISTICAL ANALYSIS

REPRODUCIBLE

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Ronald A. Fisher(1890-1962)

"TO CONSULT THE STATISTICIAN AFTER AN EXPERIMENT IS FINISHED IS OFTEN MERELY TO ASK HIM TO CONDUCT A POST MORTEM EXAMINATION. HE CAN PERHAPS SAY WHAT THE EXPERIMENT DIED OF." (1938)

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Aspects of Experimental Design

EXPERIMENTAL FACTORS

VARIABILITY

– Sources of Variance

– Replicates

BIAS

– Confounding factors

– Randomisation wherever a decision is to be made

  • Controls for both measured and unmeasured factors

– Controls

Experimental Factors

Experimental Factors

FACTORS: ASPECTS OF EXPERIMENT THAT CHANGE AND INFLUENCE THE OUTCOME OF THE EXPERIMENT

- e.g. time, weight, drug, gender, ethnicity, country, plate, cage etc.

VARIABLE TYPE DEPENDS ON TYPE OF MEASUREMENT:

- Categorical (**nominal**) , e.g. gender
- Categorical with ordering (**ordinal**), e.g. tumour grade
- **Discrete**, e.g. shoe size, number of cells
- **Continuous**, e.g. body weight in kg, height in cm

INDEPENDENT AND DEPENDENT VARIABLES

- Independent variable (IV): what you change
- Dependent variable (DV): what changes due to IV
- "**If** (**independent** variable), **then** (**dependent** variable)"

Sources of Variation

BIOLOGICAL "NOISE"

– Biological processes are inherently stochastic

– Single cells, cell populations, individuals, organs, species….

– Timepoints, cell cycle, synchronized vs. unsynchronized

TECHNICAL NOISE

– Reagents, antibodies, temperatures, pollution

– Platforms, runs, operators

CONSIDER IN ADVANCE AND CONTROL

REPLICATION REQUIRED TO CAPTURE VARIANCE

Types of Replication

BIOLOGICAL REPLICATION:

- *In vivo:*
  - Patients
  - Mice
- *In vitro:*
  - Different cell lines
  - Re-growing cells (passages)

TECHNICAL REPLICATION:

- Experimental protocol
- Measurement platform (i.e. sequencer)

PCA: Condition
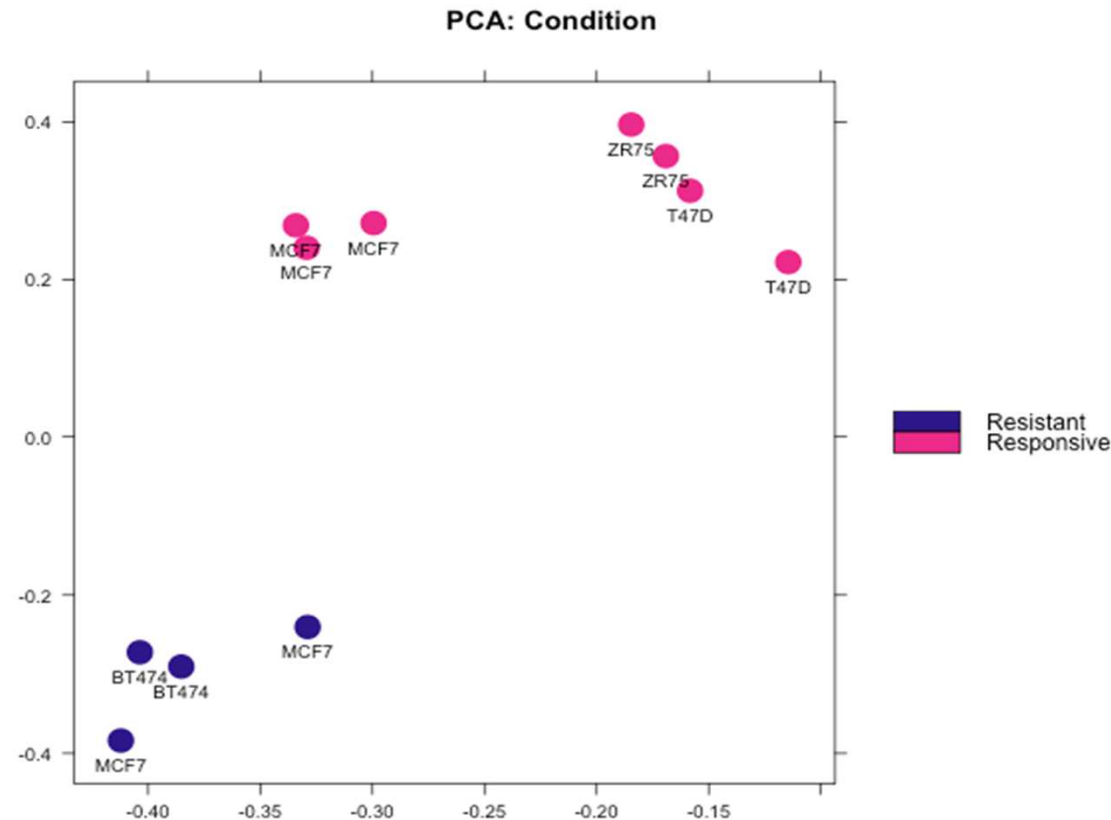
Resistant
Responsive

How many samples?

WHY DO YOU NEED REPLICATES?

CALCULATING APPROPRIATE SAMPLE SIZES

- Power calculations
- Planning for precision
- Resource equation

Effect size vs. power for unpaired t-test

sample size = 6
sd = 10
alpha = 0.05
alternative = 2-sided

- Power: the **probability** of detecting an **effect** of a specified size if present.
  - Identify and control the **sources of variability**
    - Biological variability
    - Technical variability
  - Using **appropriate numbers** of samples (sample size/replicates)
  - Power calculations estimate sample size required to detect an effect *if degree of variability is known*
    - Depends on δ, n, sd, α, H_A
  - If adding samples increases variability, that alone won't add power!
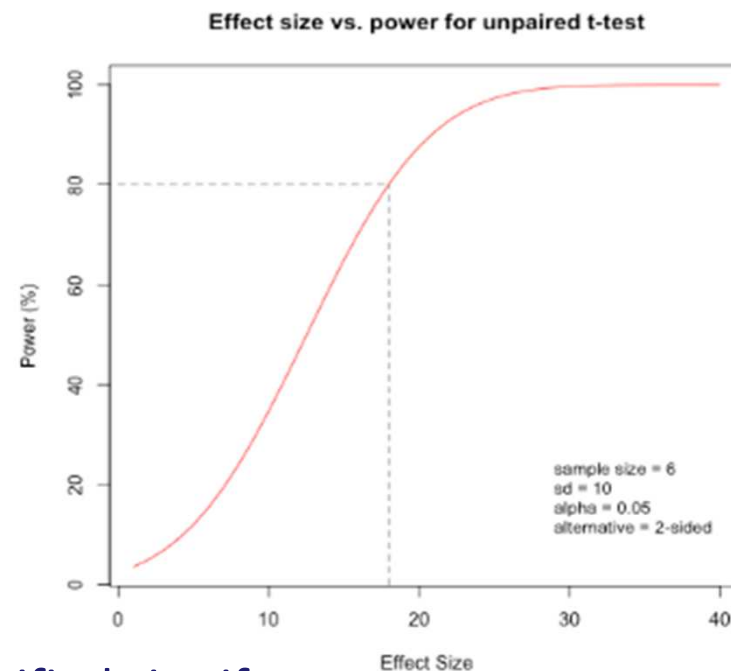
SV1   Suggest replacing this with the next slide so that we aren't emphasazing power, but happy to stick with this one if you prefer but suggest changing slide as microarray power calculations are confusing.
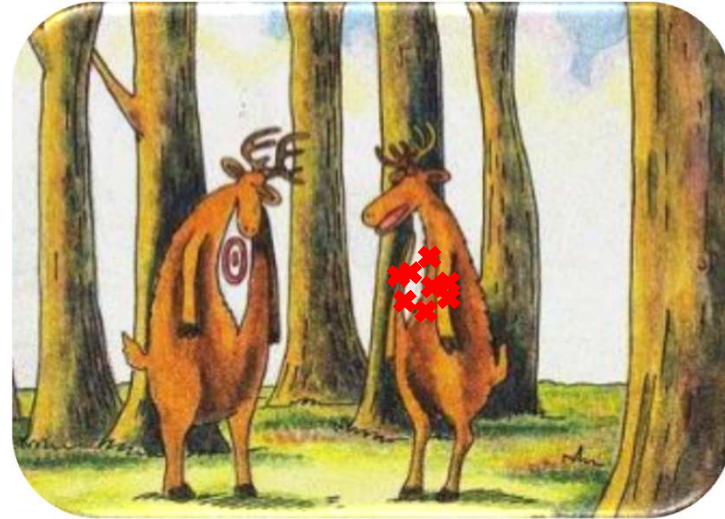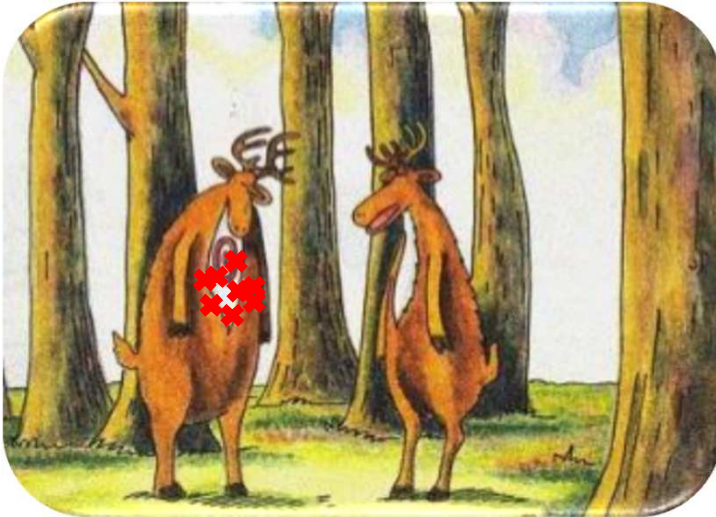
Sarah Vowler, 10/15/2014

Confounding Factors and Bias
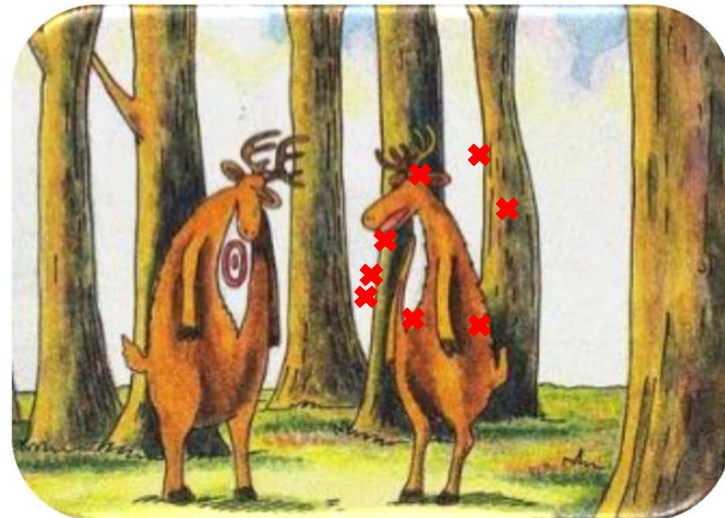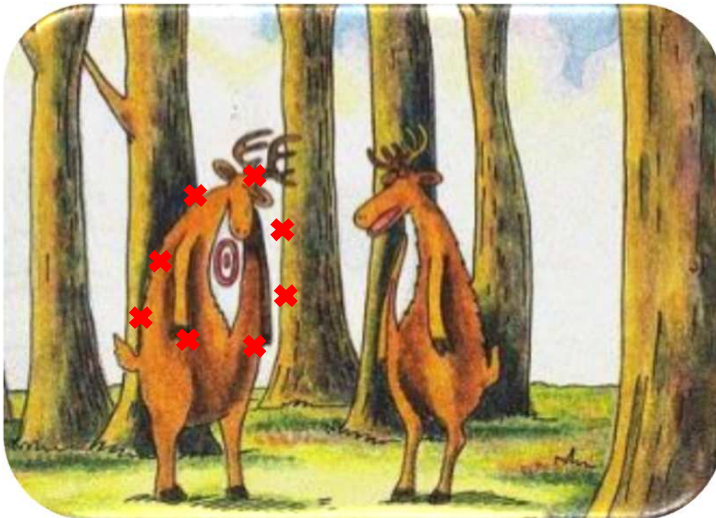
Precision, Accuracy & Bias
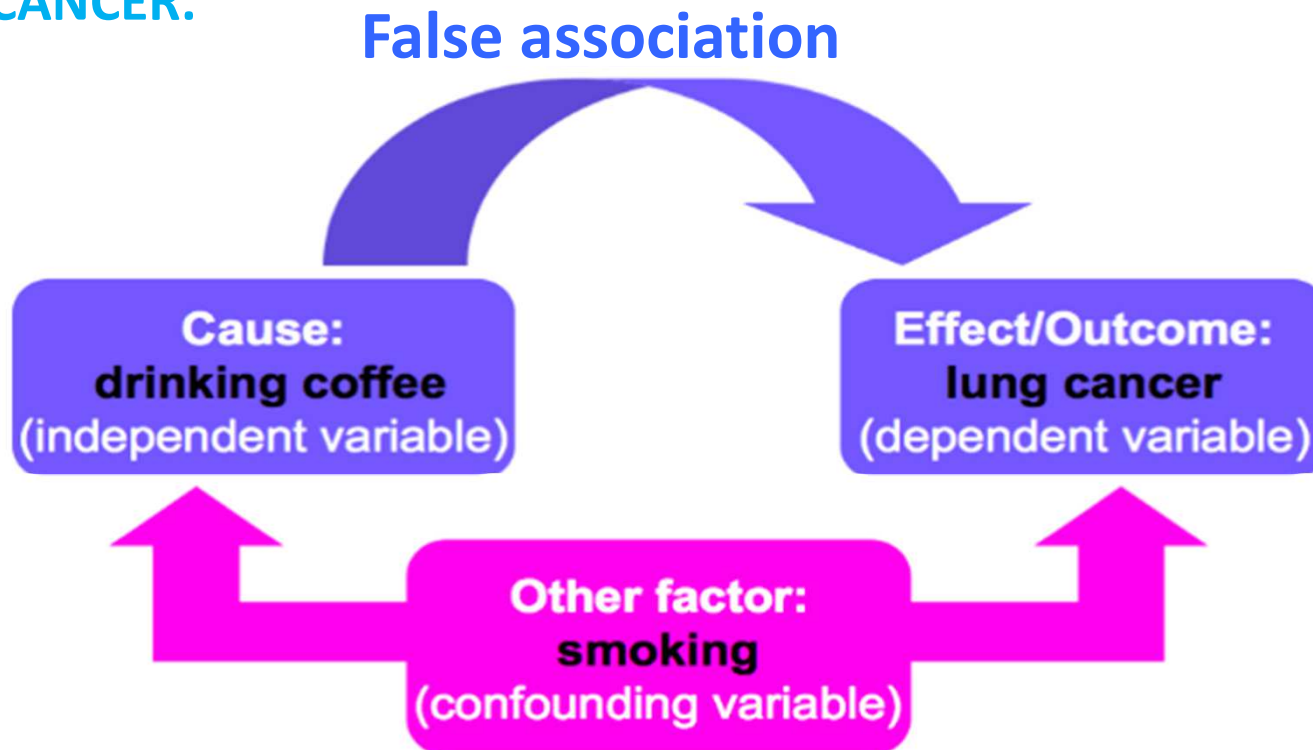
Confounding Factors

ALSO KNOWN AS EXTRANEOUS, HIDDEN, LURKING OR MASKING FACTORS, OR THE THIRD VARIABLE OR MEDIATOR VARIABLE.

MAY MASK AN ACTUAL ASSOCIATION OR FALSELY DEMONSTRATE AN APPARENT ASSOCIATION BETWEEN THE INDEPENDENT & DEPENDENT VARIABLES.

HYPOTHETICAL EXAMPLE WOULD BE A STUDY OF COFFEE DRINKING AND LUNG CANCER.

False association

Cause:
drinking coffee
(independent variable)

Effect/Outcome:
lung cancer
(dependent variable)

Other factor:
smoking
(confounding variable)

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Confounding Factors

OTHER EXAMPLES:

- Democrats were less satisfied with their sex lives than Republicans.
  (ABC poll report).

- Slightly overweight people live longer than thin people
  (US Centre for Disease Control).

INADEQUATE MANAGEMENT AND MONITORING OF CONFOUNDING FACTORS

- one of the most common causes of researchers wrongly assuming that a correlation leads to a causality.

IF A STUDY DOES NOT CONSIDER CONFOUNDING FACTORS, DON'T BELIEVE IT!

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Scienceexpress   **Report**

Genetic Signatures of Exceptional Longevity in Humans

Paola Sebastiani,[1]* Nadia Solovieff,[1] Annibale Puca,[2] Stephen W. Hartley,[1] Efthymia Melista,[3] Stacy Andersen,[4] Daniel A. Dworkis,[3] Jemma B. Wilk,[5] Richard H. Myers,[5] Martin H. Steinberg,[6] Monty Montano,[3] Clinton T. Baldwin,[6,7] Thomas T. Perls[4]*

[1]Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA. [2]IRCCS Multimedica, Milano, Italy; Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Segrate, 20122, Italy. [3]Department of Medicine, Boston University School of Medicine, Boston, MA 02118, USA. [4]Section of Geriatrics, Department of Medicine, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [5]Department of Neurology, Boston University School of Medicine, Boston, MA 02118, USA. [6]Departments of Medicine and Pediatrics, Boston University School of Medicine and Boston Medical Center, Boston, MA 02118, USA. [7]Center for Human Genetics, Boston University School of Medicine, Boston, MA 02118, USA.

- **GWAS STUDY: 800 CENTENARIANS VS. CONTROLS**
- **FOUND 150 SNPS PREDICTING CENTENARIANS WITH 77 % ACCURACY**
- **PROBLEM: THEY USED DIFFERENT SNP CHIPS FOR CENTENARIANS AND CONTROLS**
- **RETRACTED IN 2011 FOLLOWING INDEPENDENT REVIEW AND QC OF DATA**

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

http://www.the-scientist.com/blog/display/57558/
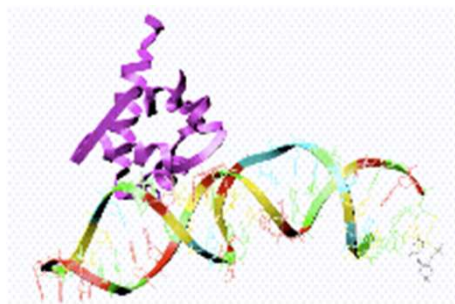
Solutions

RANDOMISATION

- Statistical analysis assume randomised comparisons
- May not see issued caused by non-randomised comparisons
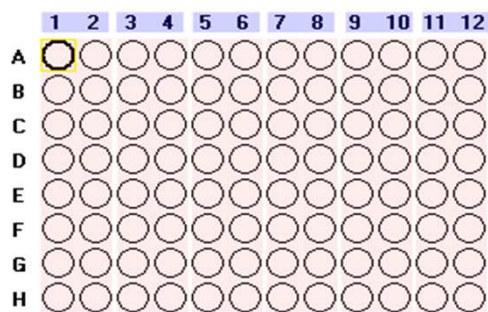- Make every decision random not arbitrary

BLINDING

- Especially important where subjective measurements are taken
- Every experiment should reach its potential degree of blinding
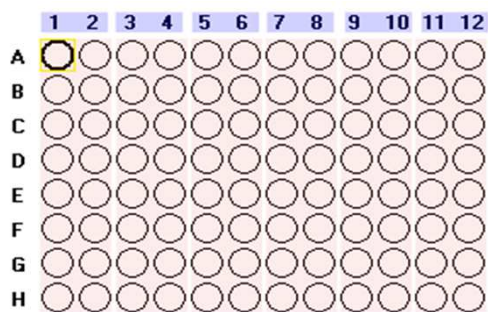
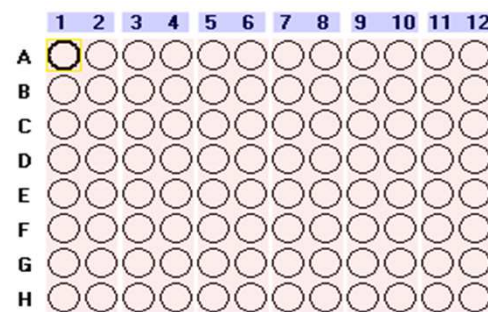# Technical Confounding Factors: Batch Effects

RNA Extraction

Day1, Plate 1

Day2, Plate 2

Day3, Plate 3

Control

Treatment 1

Treatment 2

The difference between Control, Treatment 1 and Treatment 2 is confounded by **day** and **plate**.

Solutions

RANDOMISATION

- Statistical analysis assume randomised comparisons
- May not see issues caused by non-randomised comparisons
- Make every decision random not arbitrary

BLINDING

- Especially important where subjective measurements are taken
- Every experiment should reach its potential degree of blinding

Randomised Block Design

Blocking is the arranging of *experimental units* in groups (blocks) that are similar to one another.
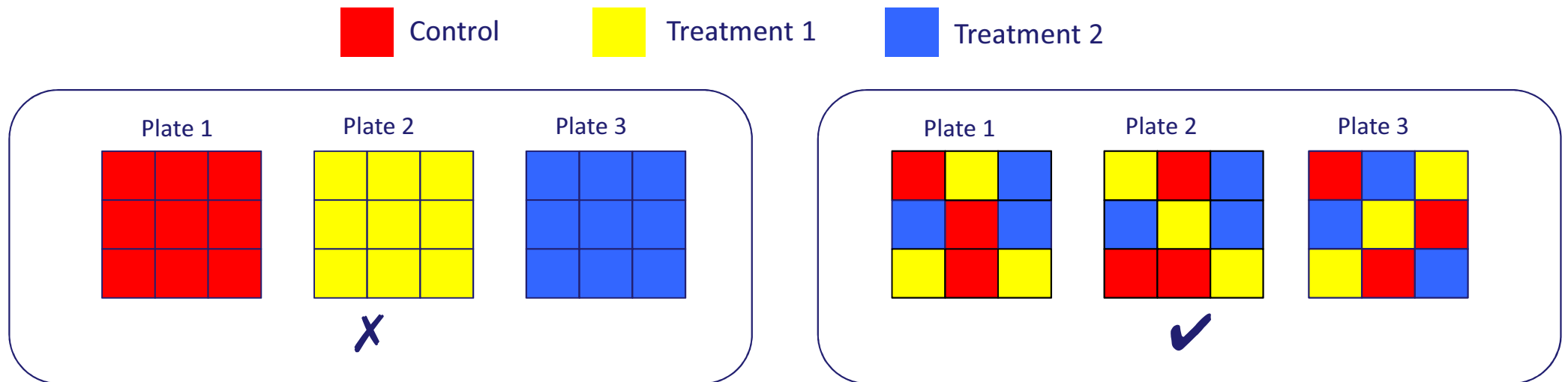
RBD across plates so that each plate contains spatially randomised **equal proportions** of:

- Control
- Treatment 1
- Treatment 2

controlling plate effects.

Randomised Block Design

Good design example: Alzheimer's study from GlaxoSmithKline

Plate effects by _plate_

Left PCA plot show _large plate effects_.
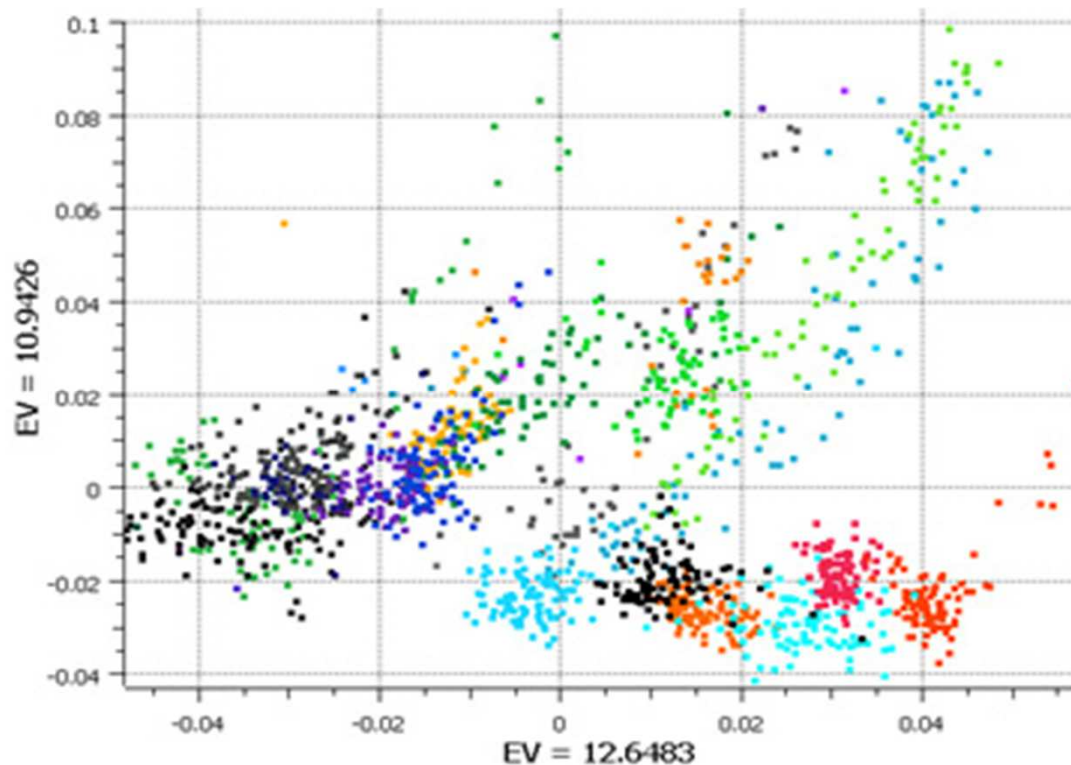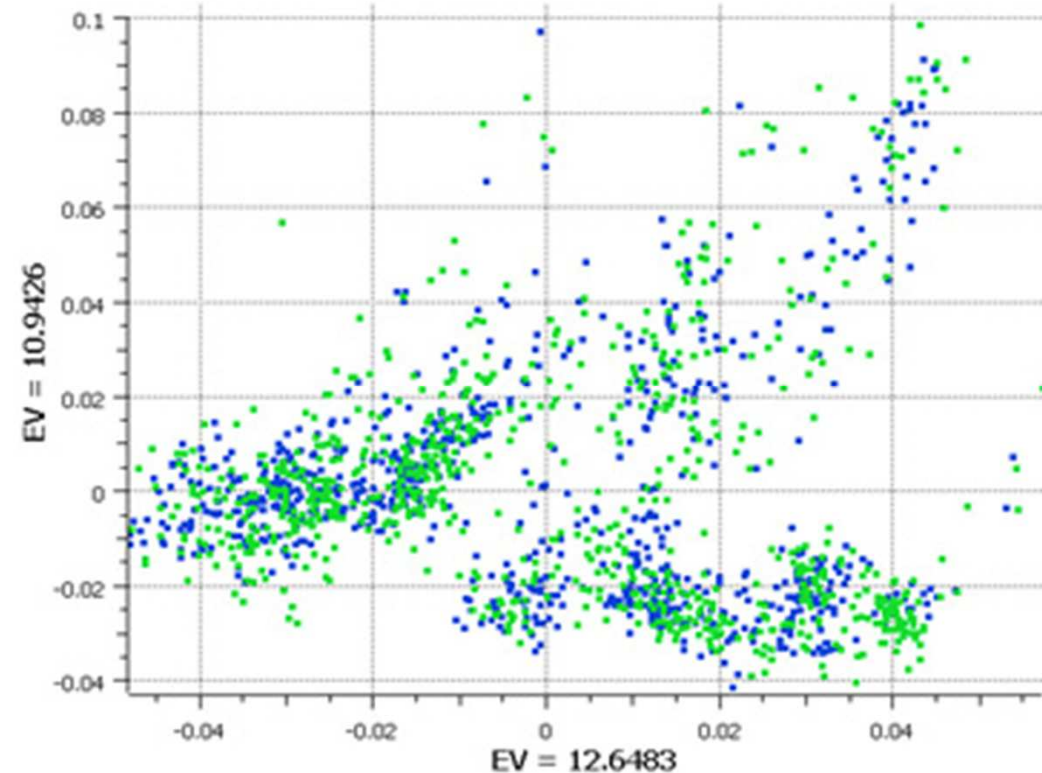Each colour corresponds to a different plate

Plate effects by _case/control_

Right PCA plot shows each plate cluster
contains _equal proportions_ of
cases (blue) and controls (green).

Experimental Controls

CONTROLLING ERRORS

- Type I: FP
  - Negative controls: should have minimal or no effect
- Type II: FN
  - Positive controls: known effect

TECHNICAL CONTROLS

- Detect/correct technical biases
- Normalise measurements (quantification)

Examples of Experimental Controls

WILD-TYPE ORGANISM (KNOCKOUTS)

INACTIVE SIRNA (SILENCING)

VEHICLE (TREATMENTS)

INPUT: FRAGMENTED CHROMATIN (CHIP)

SPIKE-INS (QUANTIFICATION/NORMALISATION)

"GOLD STANDARD" DATAPOINTS

MULTI-LEVEL CONTROLS

- e.g. contrast Vehicle/Input vs. Treatment/Input

Design Parameters for Sequencing Experiments

Design Issues: Sequencing Experiments

PLATFORMS

LIBRARY PREPS

MULTIPLEXING AND POOLING STRATEGIES

SINGLE-END VS PAIRED END

SEQUENCING DEPTH

- – Coverage
- – Lanes

VALIDATION

- – Knock-downs
- – Pull-downs

Experimental Design process at CRUK-CI

Establishing an experimental design process

– Students required to take (this) Experimental Design class

– All sequencing and proteomics experiments require experimental design review meeting

  • Simple form: EDM Form.docx

  • Attended by Scientists, Genomics/Proteomics Core, Bioinformatics Core, Statistician

  • Project opened in LIMS afterwards

– Randomisation and Layouts

  • Checkpoint for experiment

  • Project cleared for sample submission

– Keys:

  • Form and meeting not onerous

  • (Currently) not chargeable

  • Scientists agree process improves experiments!

CRI Experimental Design Meetings

TUESDAY 30 MIN SLOTS (2:00-3:00PM) WITH BIOINFORMATICS & GENOMICS/PROTEOMICS CORES

DISCUSSION:

- Planning, time-scale, cost, aims, scope, questions
- Choosing the correct technology
- Technical issues e.g. what sequencing depth?
- Sample collection and processing methods
- Sample information (meta-data) collection
- Randomisation, Blocking and Replication issues
- Analyst?
- Pilot study?
- Effect size & Sample-size calculation?

Practical: Investigation into the effect of RARα on transcription in breast cancer tissue treated with estrogen

- **RARα is a transcription factor that appears to interact with estrogen (E2) in ER+ breast cancer.**

- **We are interested in characterising this interaction by looking at how gene expression changes in breast cancer cells treated with estrogen when RARα is not present (using a siRNA in cultured cells).**

- **We wish to identify which estrogen- induced and estrogen-repressed genes are impacted by the presence or absence of RARα, and to analyse the key pathways involved.**

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Experimental Design Practical Questions I

1. **What are your objectives?**

2. **What are you measuring?**

3. **What are your primary sample groups of interest?**

4. **What controls will you use each type of sample group?**

5. **What constitutes a replicate in this experiment? Are they biological or technical? How many samples/replicates should be collected?**

6. **Sketch out the design as a matrix, with sample numbers**

7. **What sample group comparisons (contrasts) will you make with the data? Which gene set(s) will you use for pathway analysis?**

8. **What are possible confounding factors and sources of bias?**

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Experimental Design Practical Questions I

9. **How will you confirm effective silencing?**

10. **What information about your experiment should be recorded to help identify any problems should there be any?**

11. **Will you be multiplexing samples? How will you assign barcodes? Will you use pooled libraries? How many pools? How will samples be assigned to pools?**

12. **What are the sequencing parameters you need to be aware of (e.g. sequencing type and depth)?**

13. **What other types of data might be useful to assay, and how might the sequencing parameters need to change to accommodate this?**

14. **Can you think of any other design related issues that could/should be addressed?**