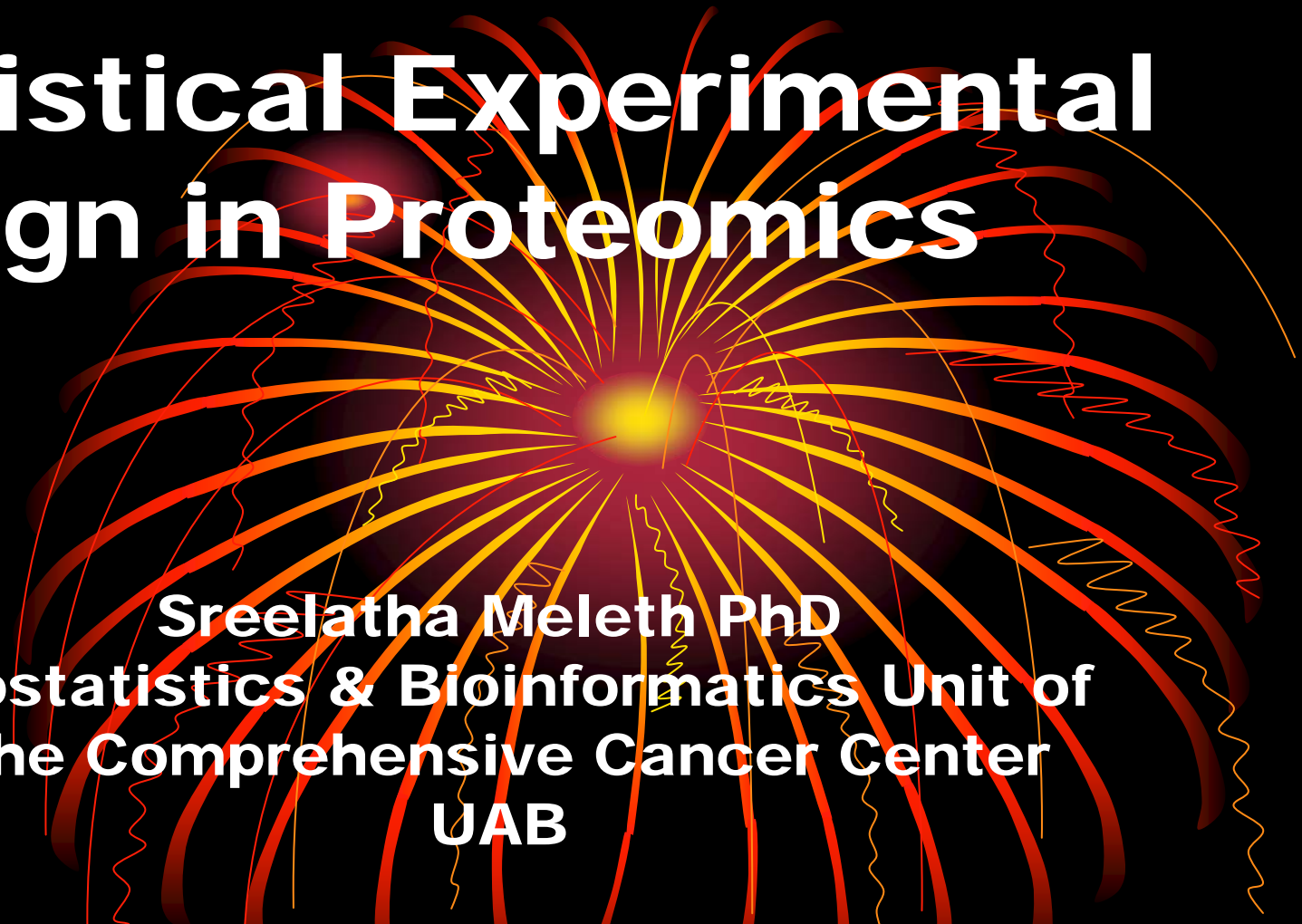


Statistical Experimental Design in Proteomics



Sreelatha Meleth PhD
Biostatistics & Bioinformatics Unit of
The Comprehensive Cancer Center
UAB

Introduction to Proteomics
2006 Workshop
University of Alabama in Birmingham

Proteomics & Biomarker Discovery Promises not kept ?



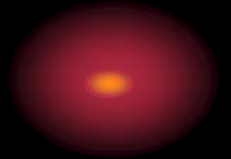
- To date the contribution of proteomics methods based on mass Spectrometry to the diagnostic armamentarium has been disappointing - Rifai N, Gillette MA, & Carr SA(2006)
- 'Concurrent with the explosion in the number of publications reporting biomarker discovery by profiling technologies such as proteomics and pattern recognition, has been the increase in evidence highlighting the susceptibility of these approaches to analytical & experimental bias' Teahan et al (2006)

Promises not kept & Statistics –

Is there a connection?



- Expensive Technologies
- Small samples
- Large number of variables – hi-dimensions
- Lack of experimental design
- Particularly – no replication, no randomization



The Statistics Connection - cont



- *Teahan et al (2006) report that*
 - Blood samples being collected onto ice vs in absence of ice
 - Over a series of serum
 - Clot contact times
 - The stability of NMR samples over time
 - Effect of freezing on the metabolic profile
- *All caused slight alterations to the NMR profile that could produce a systematic bias.*
- *Statistical Experimental Design can help one reduce or mitigate the effects of different sources of variability*

Statistical Experimental Design



- Measuring variability and attributing variability to different sources is a major part of statistical analysis
- Statistical Experimental design – aims to estimate, isolate or neutralize the variability
- Uses - Replication, Randomization & Blocking

Replication



- Biological replicates-sample size-power to detect between group variance
- Technical (same sample) replicates- helps estimate within group variance
- In techniques such as 2D gel, & microarray technical replicates also help as a quality control measure
- E.g. Are protein spots seen in all replicates of a sample?

Blocking

- Blocking - Create blocks of observations that have very similar variance
- Have every treatment group represented in each block
- e.g., Processing a 2D gel extraneous variability caused by day of processing and / technician involved
- Technicians, day will both be used as blocking factors



Randomization



- After getting a good understanding of process, and variables decide
 - Which variables to block for
 - Which variables are uncontrollable
- Uncontrollable variables neutralized by randomizing across those variables

PI / Statistician interaction



- A number of different designs-
CRBD, Latin squares, Split-plots
- Choice depends on close
consultation with PI, lab
personnel
- Is this design practical?
- You need to say 'yes it is', or
'no it is not'
- Good idea to let statistician to
see process in lab

Importance of interaction

- Experimental design has to be adapted to technology and laboratory procedures
- Case in point
 - Recent observation of DIGE gel creation
 - Process takes 4 days at a minimum
 - Monday / Tuesday start → 2nd Dim done – Thursday / Friday
 - Later start → freezing of 1st dimension gels



Importance of interaction - 2



- Initial reaction – randomize so that all groups have equal probability of being early in the week or later in the week - so that any variability caused by freezing after randomization is spread across all groups, technical and biological replicates
- Later – Is it possible to avoid this variability all together? Is it really an uncontrollable source of variation?

Benefits of Interaction - Customizing Experimental Design to suit specific technology



In 2D experiment

- Image analysis – a crucial part in the data collection
- Image Analysis is very dependent on matching of gels first within a group and then across a group
- Random assignment of treatment groups across days, technicians etc, might convert variability into random error
- However, it might also increase the variability between gels with the same treatment group to unacceptable levels.
- The need to match gels → more important to reduce variability than to distribute it as random error across the group

PS: We do not know the effect of freezing after the first dimension on spot location / intensity - need to study

Statistician /PI interaction - 3



- Allow for a learning period.. For both statisticians and you
- E.G: Randomization in a 2D gel experiment
- As above, randomization is used to convert uncontrollable sources of variation, into random error that is distributed with equal probability across all groups of interest
- Advantage – this uncontrollable variability does not affect one group disproportionately and will not be treated as an artifactual treatment effect

Statistical Analysis



- Important to use a statistician
- Most software provides two sample t-tests/ ANOVA
- Both tests above assume equality of variance and normal distribution of samples
- No provision in software to transform data or assess adherence to assumptions
- Meleth et al (2005) demonstrated that different techniques of normalization, transformation, missing data imputation alter conclusions drawn.
- Karp et al demonstrate that treating replicates as independent versus, nested alters the list of significant proteins

In summary



- Before new technologies are implemented in the search for new biomarkers it is important for the Biologist and Statistician to understand the controllable and uncontrollable sources of variation in the process
- Frequent interaction between PI, Statistician and lab personnel important to design experiments that are both practical and scientifically sound
- Data from Proteomics experiments should be sent to a statistician for analysis in order to ensure the validity of the results.

Questions?



Thank you!

