



CALGB Bioinformatics Unit

Jennifer Shoemaker, Ph.D.
CALGB Statistical Center

CALGB Group Meeting, June 2005

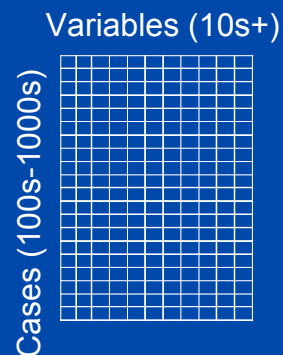
Mission Statement

To collaborate in the design and analysis
of studies with a bioinformatics
component.

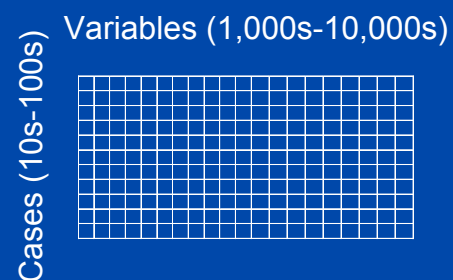
Definition of Bioinformatics

“The collection, analysis, and integration of biological (often genomic, biochemical, or molecular) information using techniques in computer science, statistics, and mathematics.”

Typical Clinical Study



Typical Gene Expression Study



*Figure 1. Adapted from Kohane et al. 2002. Microarrays for an Integrative Genomics. MIT Press (Cambridge, MA).

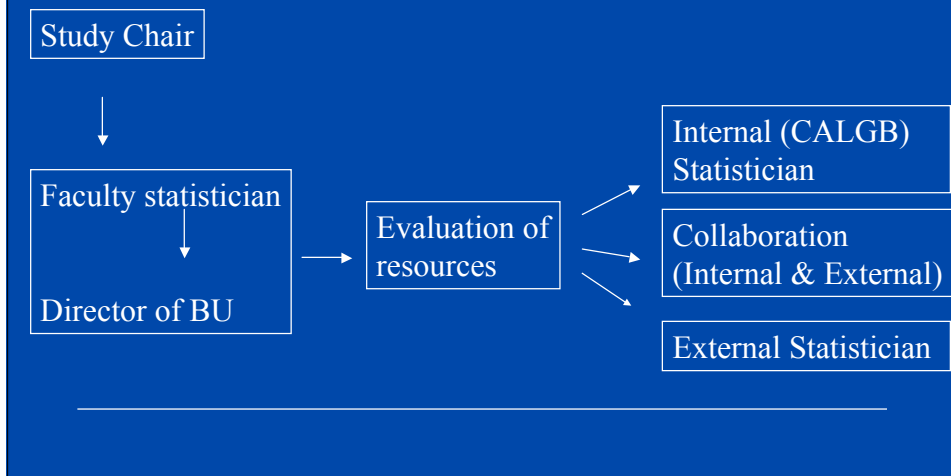
CALGB and Bioinformatics

- Natural extension of correlative sciences
 - Central resource
 - Data storage
 - Stewardship, QC, Standardization
 - Methodological tools
 - Volume of studies
 - Estimate minimum of 25 studies
-

Review

If the concept includes a bioinformatics component, it must be reviewed by the Bioinformatics unit before being submitted to the Executive Committee.

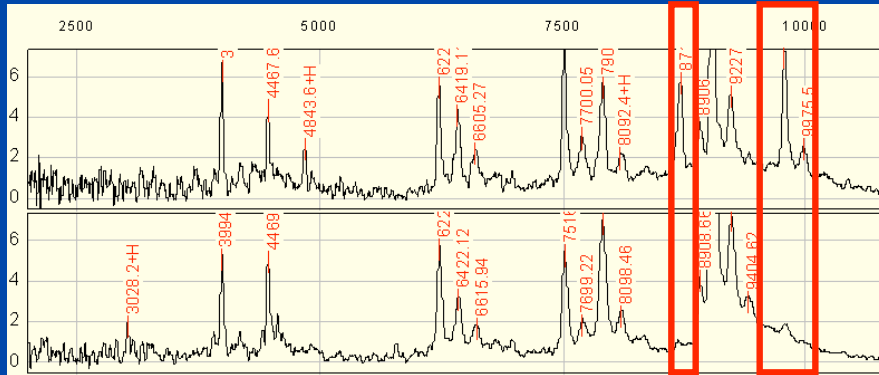
Procedure for Adding a Bioinformatics Component



Methods for Handling Bioinformatics Components

- **Internal (CALGB) Statistician**
 - 90203 (Microarray, SNPs, Proteomics)
 - 80405 (Proteomics)
 - 9665 Leukemia Tissue Bank (Microarray)
- **Collaboration**
 - 150201 (Proteomics)
 - 9665 Leukemia Tissue Bank (Microarray, Restriction Landmark Genome Scanning)
- **External**
 - 159806 (Microarray)
 - 50303 (Microarray)
 - 9665 Leukemia Tissue Bank (Microarray)
 - 150303 (Microarray)

Proteomics Profiling Studies



Courtesy: CIPHERgen

Proteomics Raw Data

- 1.5 MBs per spectrum
- Small study:
 - 30 samples
 - 10 fractions per sample
 - 10 runs per fraction

Name	Size	Type	Date Modified
01010001-10.23.03.txt	1,530 KB	Text Document	3/31/2004 1:16 PM
01010002-10.23.03.txt	1,525 KB	Text Document	3/31/2004 1:16 PM
01010003-10.23.03.txt	1,527 KB	Text Document	3/31/2004 1:16 PM
01010004-10.23.03.txt	1,521 KB	Text Document	3/31/2004 1:17 PM
01010005-10.23.03.txt	1,544 KB	Text Document	3/31/2004 1:17 PM
02020001-03.03.04.txt			
02020002-03.03.04.txt			
02020003-03.03.04.txt			
02020004-03.03.04.txt			
02020005-03.03.04.txt			
03010001-10.30.03.txt	21475, 569916	70, 264	
03010002-10.30.03.txt	21476, 599857	67, 4244	
03010003-10.30.03.txt	21477, 509818	67, 4244	
03010004-10.30.03.txt	21478, 479800	76, 653	
03010005-10.30.03.txt	21479, 449805	82, 321	
03010006-10.30.03.txt	21480, 419831	75, 2332	
03010007-10.30.03.txt	21481, 389880	74, 5233	
03010008-10.30.03.txt	21482, 359950	73, 11035	
03010009-10.30.03.txt	21483, 330043	79, 4926	

$$1.5 \times 30 \times 10 \times 10 = 4.5 \text{ GB}$$

Human genome = 3 GB

7 CDs = 4.9 GB

1 DVD = 4.5 GB

Bioinformatics Analysis/Data Storage

- Supercomputing Centers:
 - Pittsburgh
 - Ohio
 - Minnesota
 - Lawrence Berkeley
 - Open Sources initiative at Duke
-

Procedure for Utilizing External Statistician

- Study Chair
 - Identify all external statisticians
 - Send external statisticians' CVs to Director of BU
 - Prepare material for Duke IRB submission
 - Director of BU
 - Review qualifications of external statisticians
 - Convey responsibilities to external statisticians
-

Responsibilities of External Statistician

- Sign CALGB agreement
 - Help in data definition
 - Write statistical section
 - Respond to reviewers' comments (EC, CTEP, CTSU, CIRB)
 - Perform data analysis
 - abstracts
 - manuscripts
-

Policy for Handling Data

Laboratories submit all molecular data to CALGB Statistical Center.

Procedure for Handling Data

- Laboratories submit all molecular data to CALGB Statistical Center
 - CALGB statistician merges molecular data with clinical data
 - CALGB statistician de-identifies merged data
 - CALGB statistician sends de-identified data to external statistician
-

Grant Support

- If the Study Chair is seeking additional grant funding, include support for the Statistical Center
 - Statistician(s)
 - Data Coordinator
 - Form Designer
 - IS Programmer
 - Consult with Director of BU for appropriate percentages
-

Contact Information

Jennifer Shoemaker, Ph.D.

Assistant Research Professor

Department of Biostatistics and Bioinformatics

Duke Bioinformatics Shared Resource

Director, CALGB Bioinformatics Unit

DUMC Box 3958

Durham, NC 27710 USA

Phone: 919 668-2655 **Fax:** 919 681-8028

