Lecture 2: Statistical bioinformatics for gene expression data

Analyzing Gene Expression Array Data

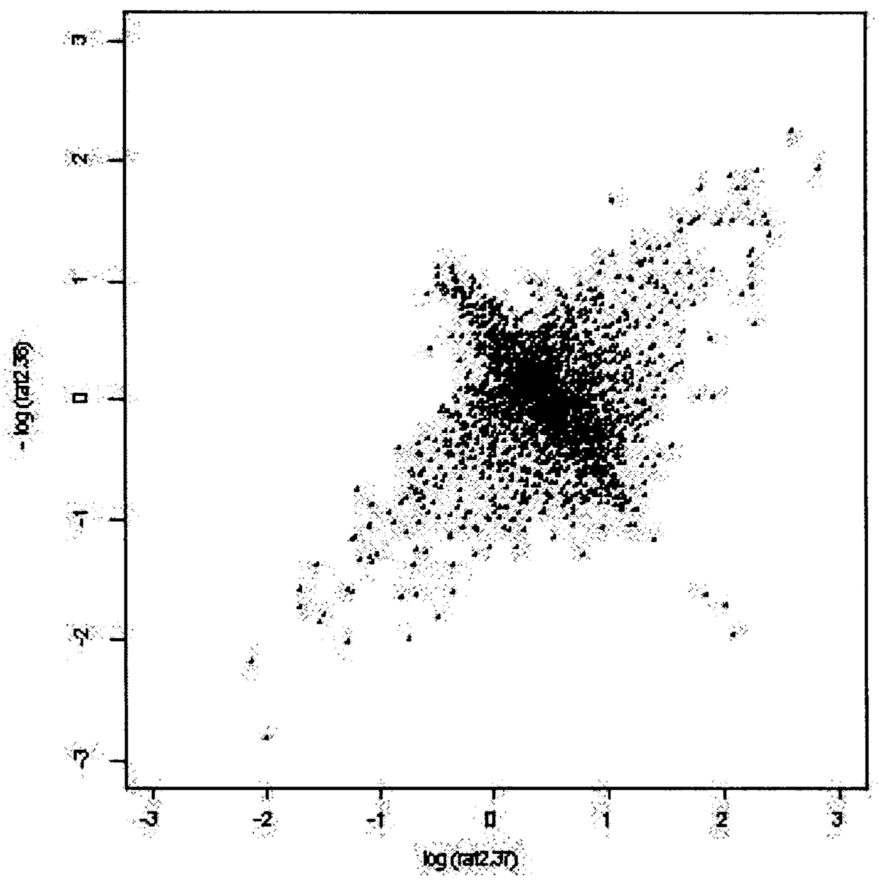
Need pre-processing and pre-screening to avoid (large number of) irrelevant results from various artifacts

- quality control and pre-processing:
    - thresholding to reduce the noise level in low intensity
    - normalization (within-chip and between-chip)
    - reproducibility
      Microarray: ratio statistics (Chen et al. 97) --> log transformation
      Affymetrix: ave. difference --> log transformation
- pre-screening and subsetting:
    - genes with major expression variations: fold-change discovery
    - genes with distinctive patterns in specific cases: clustering or classification

Myth 2: Can do without a statistical design?

- Various statistical factors of variability in Microarray
  - gene and variety (types of sample, treatment, time...)
  - individual sample, chip array, and dye (microarray)

Artifact in dye-swap array data



- Replication and Experimental design (blocking)
  - Replicates of genes on a chip
  - Statistical significance tests based on replicated chip data (MT Lee, 2000; Speed, 2000)
  - Replicated chips for treatments, especially for interaction (Kerr and Churchill, 2001, Lee et al., 2001)

Example: Experimental design on an array study

- Microarray study on comparing a treatment effect at two different time points with two individual replicates

Chip 1		Chip 2		Chip 3		Chip 4	
Cy3	Cy5	Cy3	Cy5	Cy3	Cy5	Cy3	Cy5
I1-T1	Ref	Ref	I1-T2	Ref	I2-T1	I2-T2	Ref

- Replicates for arrays, dyes, individuals are shared.

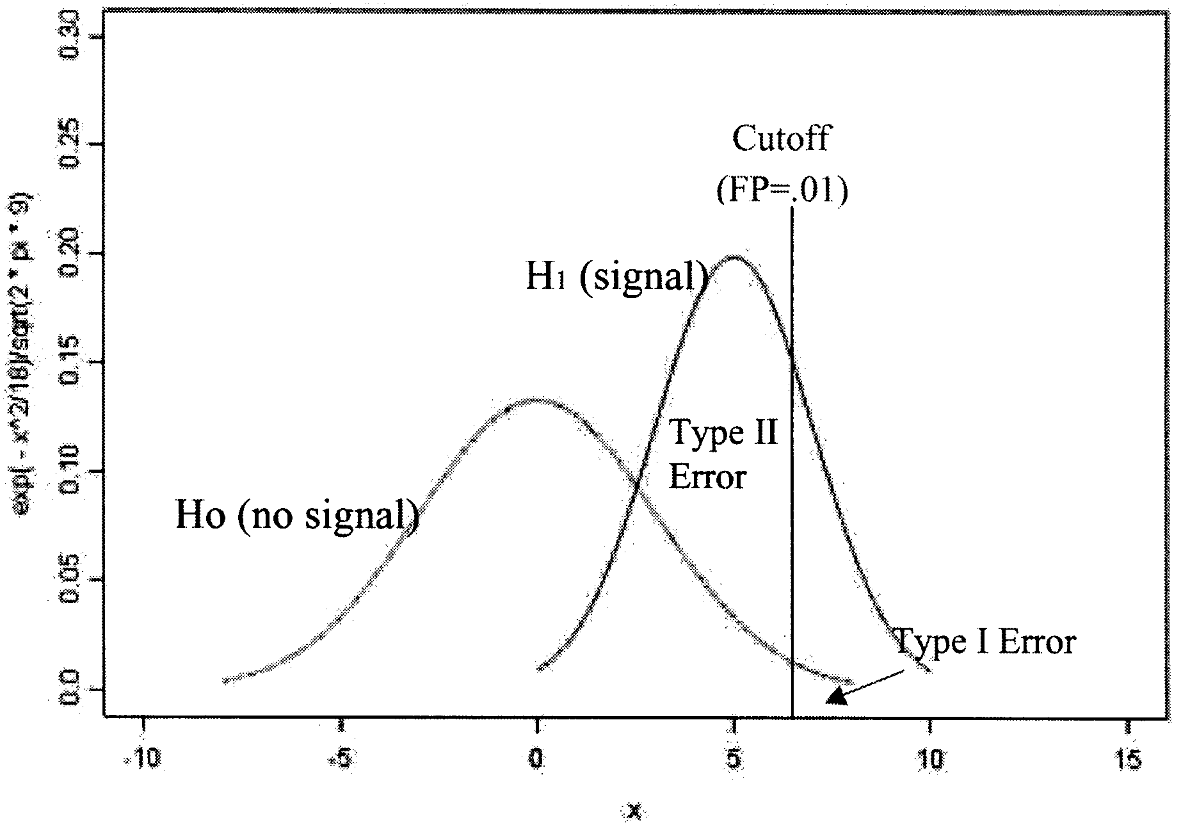
- Treatment and time point factors are separately replicated from individual, array, and dye factors.

Myth 3: Experimental confirmation, such as Northern, RT-PCR, or Western blot is the only way to validate the bioinformatic findings?

Statistical Validation

- Baseline distribution: duplicates (sometimes triplicates)
- Independent chip replicates: independent sample collections and RNA preps are most ideal to validate bioinformatic findings even within biological variability
- Limited resources: Replicates for most variable biological factors

Type I (FP) and II (FN) Errors

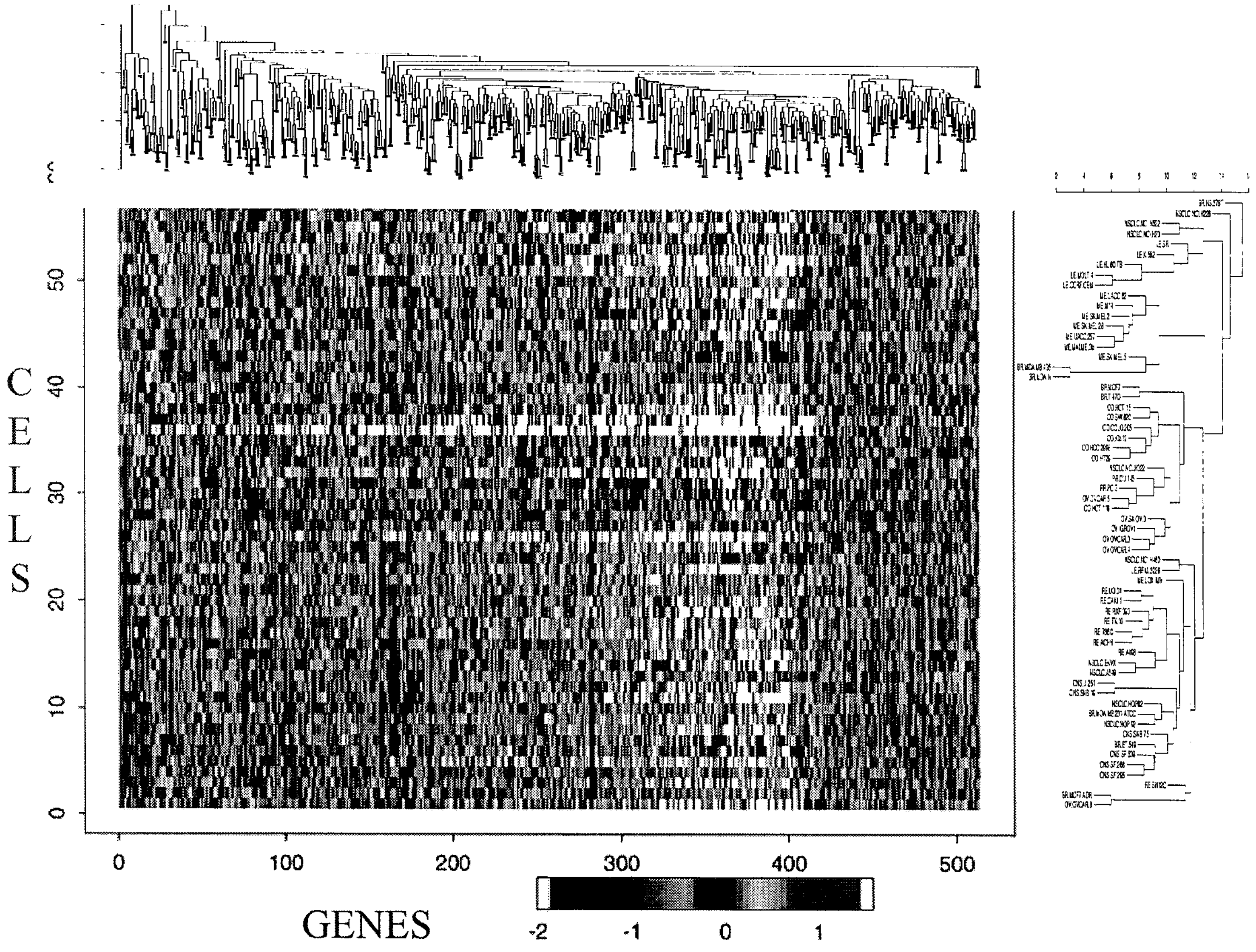


Clustering Analysis

- Searching for the groups (clusters) in the data, based on a measure or distance index of similarity or dissimilarity
- Dissimilarity distance metric:
  - $d(x, x) = 0, d(x, y) > 0, d(x, y) = d(y, x)$
  - Examples:
    - Euclidean: $d(x, y) = \{w_k(x_k - y_k)^2\}^{1/2}$
    - Manhattan: $d(x, y) = w_k |x_k - y_k|$
    - Maximum, Binary...

  - Correlation distance: $d(x, y) = 1 - r(x, y)$, $r(x, y)$ is Pearson sample correlation coefficient (Spearman, binary).
    - Relationship between Euclidean and correlation distances when $w_k = 1$:
      - If x and y are standardized, i.e., subtracted by each mean and divided by each standard deviation, then two distances are equivalent.
      - $(x_k - y_k)^2 = x_k^2 + y_k^2 - 2x_k y_k = 2(1 - x_k y_k) = 2\,(1 - r(x, y))$

Hierarchical clustering: Cluster-Image Analysis



Clustering algorithm I: Partitioning

- Partitioning: divides the data into a pre-specified number of subsets

- Kmeans (Ruspini, 1970; centroid): Iterative relocation clustering
- Pam (partitioning around medoids—k representative objects; robust than Kmeans)
- Clara (Clustering large applications; distance matrices for subsets)
- Fuzzy algorithm (fractions of membership; e.g., fanny)
- Mclust (probability-based model; Bayes factor for choosing k clusters)
- MDS (multidimensional scaling; Tamayo et al., 1999)

Clustering algorithm II: Hierarchical allocation

- Agglomerative methods
  - average linkage (group average)
  - single linkage (min, nearest)
  - complete linkage (max, furthest)
  - Example
    - hclust (hierarchical; e.g., Eisen et al, 1998, Scherf et al., 2000)
- Divisive methods
  - mona (monothetic—single variable division)
  - diana (polythetic)

Statistical classification and clsutering

- Identification of new or unknown classes (Unsupervised learning)
- Classification into known classes (Supervised learning)
- Identification of “best” predictor variables—variable selection, e.g. marker genes in microarray data

Aggregating predictors(Breiman, '98)

- Gains accuracy by aggregating predictors built from a learning set
    - Predictors are aggregated by **voting**
    - Bagging (bootstrap aggregating)

    - perturbed learning sets by bootstrap

Example 1: *Gene voting* (Golub et al., 1999)

- For binary classification, each gene casts a vote for class 1 or 2 among p samples, and the votes are aggregated over genes. For gene g_j the vote is $v_j = a_j (g_j - b_j)$, where $a_j = (s\mu_1 - s\mu_2)/(s\sigma_1 + s\sigma_2)$, $b_j = (s\mu_1 + s\mu_2)/2$.

- Classification between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL)

  - ALL (27 cases), AML (11)
  - Affymetrix chip data (6,817 genes)
  - Choose 50 informative genes based on high (binary) association with classification.
  - prediction strength (PS) = $(V_{win} - V_{lose})/(V_{win} + V_{lose})$
  - 36/38 (ps> 0.3), 2/38 (ps < 0.3) by leave-one-out cross-validation

Example 2: Hierarchical Clustering (Eisen et al., 1998)

- time-course gene expression studies
  - ➢growth model response in human cells
  - ➢cell cycles of budding yeast → Identify groups of genes by their functions and origins

Example 3: Self-organizing maps (SOM; Tamayo, et al., 1999; *GeneCluster*)

- Hematopoietic cell lines (HL60, U937, Jurkat, and NB4): 4x3 SOM

- Yeast data in Eisen et al. reanalyzed by 6x5 SOM

Supporting Vector Machine(Brown et al., 2000; Furey et al., 2000)

- Separate a given set of binary labeled training data with a hyper-plane that is maximally distant from them, known as the maximal margin hyper-plane.

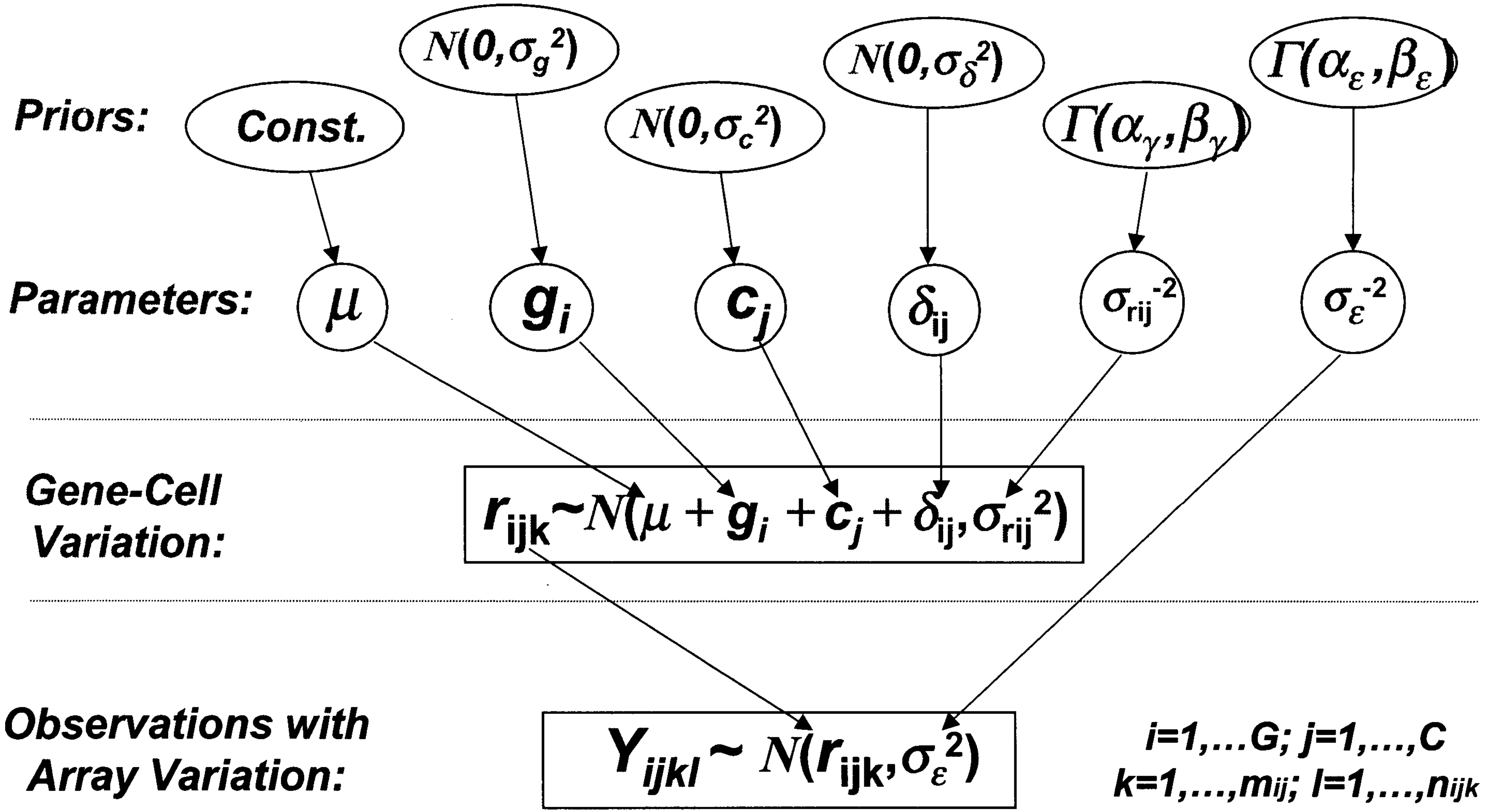
- Base on a kernel, such as a polynomial of dot products, the current data space will be embedded in a higher dimensional space.

- Using a training set, derive a hyper-plane with maximal separation and validate against a validation set.

Estimation of Gene-Cell interactive responses frommicroarray data

- $Y_{ij} - G_i - C_j = (GC)_{ij} + \varepsilon_{ij}$ (gene-cell specificity)
  or $Y_{ij} = G_i + C_j + (GC)_{ij} + \varepsilon_{ij}$ (gene-cell effects)

- Loop experimental design and ANOVA (Kerr and Churchill, 2000)
- Hierarchical error model (Lee et al., 2001)



Priors: Const. $N(0,\sigma_g{}^2)$ $N(0,\sigma_c{}^2)$ $N(0,\sigma_\delta{}^2)$ $\Gamma(\alpha_\gamma,\beta_\gamma)$ $\Gamma(\alpha_\varepsilon,\beta_\varepsilon)$

Parameters: μ g_i c_j δ_{ij} $\sigma_{rij}{}^{-2}$ $\sigma_\varepsilon{}^{-2}$

Gene-Cell Variation: $r_{ijk} \sim N(\mu + g_i + c_j + \delta_{ij}, \sigma_{rij}{}^2)$

Observations with Array Variation: $Y_{ijkl} \sim N(r_{ijk}, \sigma_\varepsilon{}^2)$

i=1,…G; j=1,…,C
k=1,…,m_{ij}; l=1,…,n_{ijk}

Summary

- Gene expression studies require statistical experimental designs and validation before laboratory confirmation.
- Various clustering approaches, such as hierarchical, Kmeans, SOM are commonly used for unsupervised learning in gene expression data.
- Several classification methods, such as gene voting, SVM, or discriminant analysis are used for supervised lerning, where well-defined response classification is possible.
- Estimating gene-condition interaction effects require advanced, computationally-intensive statistical approaches.

Jae K. Lee, Ph.D.

Division of Biostatistics and Epidemiology
Department of Health Evaluation Sciences,
University of Virginia School of Medicine
P.O.Box 800717
Charlottesville, VA 22908
(Voice) 804.982-1033, (Fax) 804.924-8437, (Email) jaeklee@virginia.edu

Education

1990-1995 *Ph.D.* (Statistics), University of Wisconsin-Madison, Madison, Wisconsin.

1985-1987 *M.Sc.* (Statistics), Seoul National University, Seoul, South Korea.

1981-1985 *B.Sc.* (Mathematics), Seoul National University, Seoul, South Korea.

Dr. Lee received his Ph.D. in Statistics from the University of Wisconsin-Madison in 1995. Before coming to the University of Virginia, he was a research associate at the Center for Computational and Experimental Genomics of the University of Southern California and a research scientist at the Bioinformatics Group of the Laboratory of Molecular Pharmacology of the National Cancer Institute (NCI) of the National Institutes of Health (NIH). Dr. Lee has extensive experience in statistical research in molecular genetics and bioinformatics. He has worked on and familiar with statistical approaches in population inference, DNA structure analysis, linkage association study for human genetic diseases, and high throughput gene chip technologies. He has also applied and developed state-of-the-art technologies of statistics and computer sciences to attack various challenging problems in

molecular biology and medicine, such as linkage association study for identifying multiple trait loci of pedigree data and anticancer gene-drug discovery on high throughput gene expression data. Dr. Lee teaches a new course Statistical Bioinformatics in Medcine (HES795) at the University of Virginia. Recent Publications are:

1. **Lee, J.K., Nordheim, E.V., and Kang, H. (1996). Inference for Lethal Gene Estimation with Application in Plants. *Biometrics*, 52, 451-462.**
2. **Lee, J.K., Lascoux, M., and Nordheim, E.V. (1996). Number of Lethal Equivalents *in* Human Populations: How good are the Previous Estimates? *Heredity*, 77, 209-216.**
3. **Lee, J.K. Assessment of Deleterious Gene Models via HPD Predictive p-values (1996). *Case Study 3: Bayesian Statistics in Science and Technology*, 387-397.**
4. **Lee, J.K., Dancik, V., and Waterman, M.S. (1998). Estimation of the Restriction Sites Observed by Optical Mapping Using Markov Chain Monte Carlo. *Journal of Computational Biology*, 5, 505-515.**
5. **Lascoux, M. and Lee, J.K. (1998). Characterization of Deleterious Loci in the Fast-Cycling B. napus L. Base Population. *Genetica*, 104, 161-170.**
6. **Tanabe L, Scherf U, Smith LH, Lee JK, Hunter L, Weinstein JN (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques*, 27(6), 1210-4, 1216-7.**
7. **Lee, J.K., Lascoux, M., Newton, M.A., and Nordheim, E.V. (1999). A Study of Deleterious Gene Structure in plants using Markov chain Monte Carlo. *Biometrics*, 55, 376-386.**
8. **Scherf, U., Ross, D.T., Waltham, M., Smith, L.H., Lee, J.K., Kohn, K.W., Reinhold, W.C., Myers, T.G., Andrews, D.T., Scudiero, D.A., Eisen, M.B., Sausville, E.A., Pommier, Y., Botstein, D., Brown, P.O., and Weinstein, J.N. (2000). A cDNA microarray gene expression database for the molecular pharmacology of cancer. *Nature Genetics*, 24 (3), 236-244.**
9. **Shi, L.M., Fan, Y., Lee, J.K., Myers, T., Waltham, M., Andrews, A., Scherf, U., Paull, K., Weinstein, J.N. Mining and Visualizing Large Anticancer Drug Databases (2000). *J. of Chem. Inf. & Com. Sci.*, 40 (2), 367-379**
10. **Woodfolk, J.A., Sung, S.J., Benjamin, D.C., Lee, J.K., Platts-Mills, T.A.E. (2000). Distinct human T cell repertoires mediate immediate and delayed-**

type hypersensitivity to the Trichophyton antigen, Tri r 2'. *J. of Immunology*, 165, 4379-4387.

11. Lee, J.K. and Thomas, D.C. (2000). Performance of Markov Chain Monte Carlo Approaches for Mapping Genes in Oligogenic Models with an Unknown Number of Loci. *American Journal of Human Genetics* 67: 1232-1250.
12. Lee, J.K., Tavaré,S., Brown, J., and Deonier, R.C. Identifying IHF-Binding Sites on DNA from Predicted DNA Structure. *Revised for Journal of Molecular Biology.*
13. Lee, J.K., Scherf, U., Smith, L.H., Tanabe, L., and Weinstein, J.N. Analyzing Genomic and Pharmacological Data in the National Cancer Insitute's Drug Discovery Program: A Bayesian Hierarchical Effects Approach Using Gibbs Sampling. *Submitted to Bioinformatics.*
14. Staunton, J.E., Slonim, D.L, Coller, H.A., Tamayo,P., Angelo, M.J., Park, J., Scherf, U., Jae K. Lee, J.K., Weinstein, J.N., Mesirov, J.P., Lander, E.S., and Golub, T.R. Chemosensitivity Prediction by Transcriptional Profiling, *submitted to PNAS.*