GS01 0163
Analysis of Microarray Data

Keith Baggerly and Kevin Coombes
Department of Bioinformatics and Computational Biology
UT M. D. Anderson Cancer Center
kabagg@mdanderson.org
kcoombes@mdanderson.org

4 December 2007

Lecture 27: Meta-Analysis of CLL Studies

- Background on CLL
- CLL on U95 Arrays
- CLL on the Lymphochip
- Research Genetics GeneFilters
- Meta-Analysis
- Predicting Mutation Status

Background on CLL

Chronic lymphocytic leukemia (CLL) is the most prevalent leukemia in the western world, accounting for 25% of all leukemias in the US.

The median survival is about 9 years, but may be as short as 1 or 2 years even in early stage patients.

Traditional prognostic factors (including clinical stage, gender, pattern of bone marrow involvement, lymphocyte doubling time, and serum β_2 microglobulin) do not adequately account for the heterogeneity in outcome.

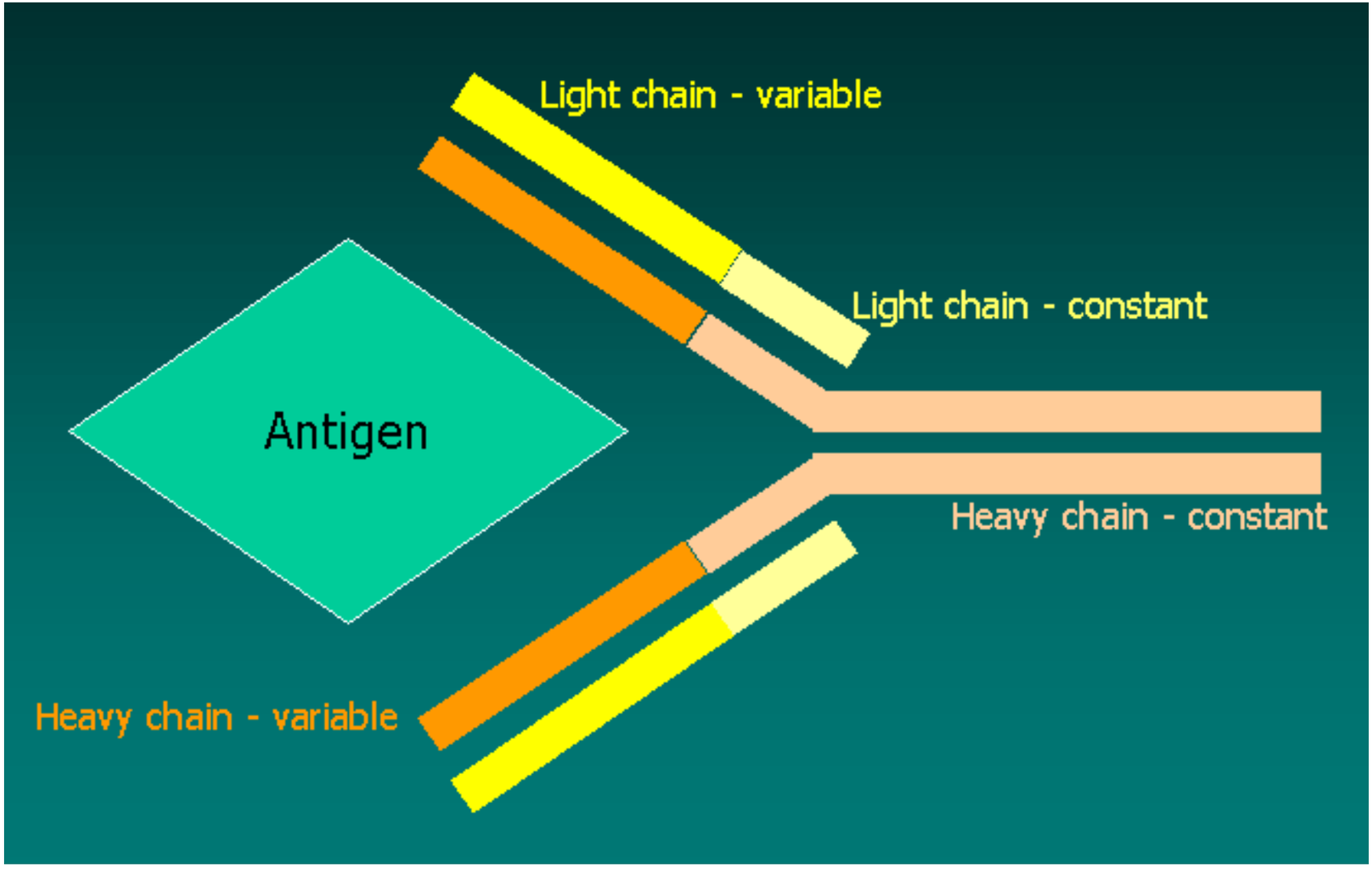
Somatic hypermutation status

One of the most promising new markers of prognosis in CLL is the somatic hypermutation (SHM) status of the immunoglobulin heavy chain variable region (IgV_H) genes.

When normal B cells are exposed to antigens, they migrate to the germinal center (GC) of the lymph nodes where the IgV_H genes are mutated. This process is believed to contribute to the immune system's ability to respond adaptively to pathogens.

- About 40% of CLL patients have unmutated IgV_H genes and a poor prognosis (median survival: 8 years);
- The other 60% have mutated IgV_H genes and a better prognosis (median survival: 25 years).

The Structure of an Antibody



Combinatorial Antibody Diversity

- Kappa Light Chain
  - Chromosome 2p11
  - 1 constant region (IGKC)
  - 5 joining regions (IGKJ)
  - 31–35 variable regions (IGKV)
- Lambda light chain
  - Chromosome 22q11
  - 4–5 constant regions (IGLC)
  - 4–5 joining regions (IGLJ)
  - 29–33 variable regions (IGLV)
- Total of ~ 160 light chains
- Heavy chain
  - Chromosome 14q32
  - 9 constant regions (IGHC)
  - 23 diversity regions (IGHD)
  - 6 joining regions (IGHJ)
  - 38–46 variable regions (IGHV)
- Total of $\sim 6,000$ heavy chains
- Roughly $1,000,000$ different antibodies achieved by combinatorics

Somatic Hypermutation (SHM)

Additional antiboidy diversity is provided by the deliberate introduction of mutations into the immunoglobulin genes. After B cells are exposed to antigen, they mature in the lymph nodes, where the IGHV gene undergoes SHM.

- IGVH region spans about 1.3 Mb on chr. 14.
- Individual IGHV genes are 300–500 bases long.
- Mustdetermine which IGVH gene is being used, sequence full length, and compare it to germline
  - Mutated: homology $\leq 98\%$
  - Unmutated: homology $> 98\%$
- Sequencing is difficult and expensive

CLL on U95A Arrays

Reference: Klein U, et al. (2001) *Gene expression profiling of B cell chronic lymphocytic leukemia reveals a homogeneous phenotype related to memory B cells.* J Exp Med **194**: 1625–1638.

This study, from Riccardo Dalla-Favera's laboratory at Columbia, used Affymetrix U95A GeneChips to try to understand the subtypes of CLL.

We'll start by reviewing the results they reported.

Samples studied on U95A

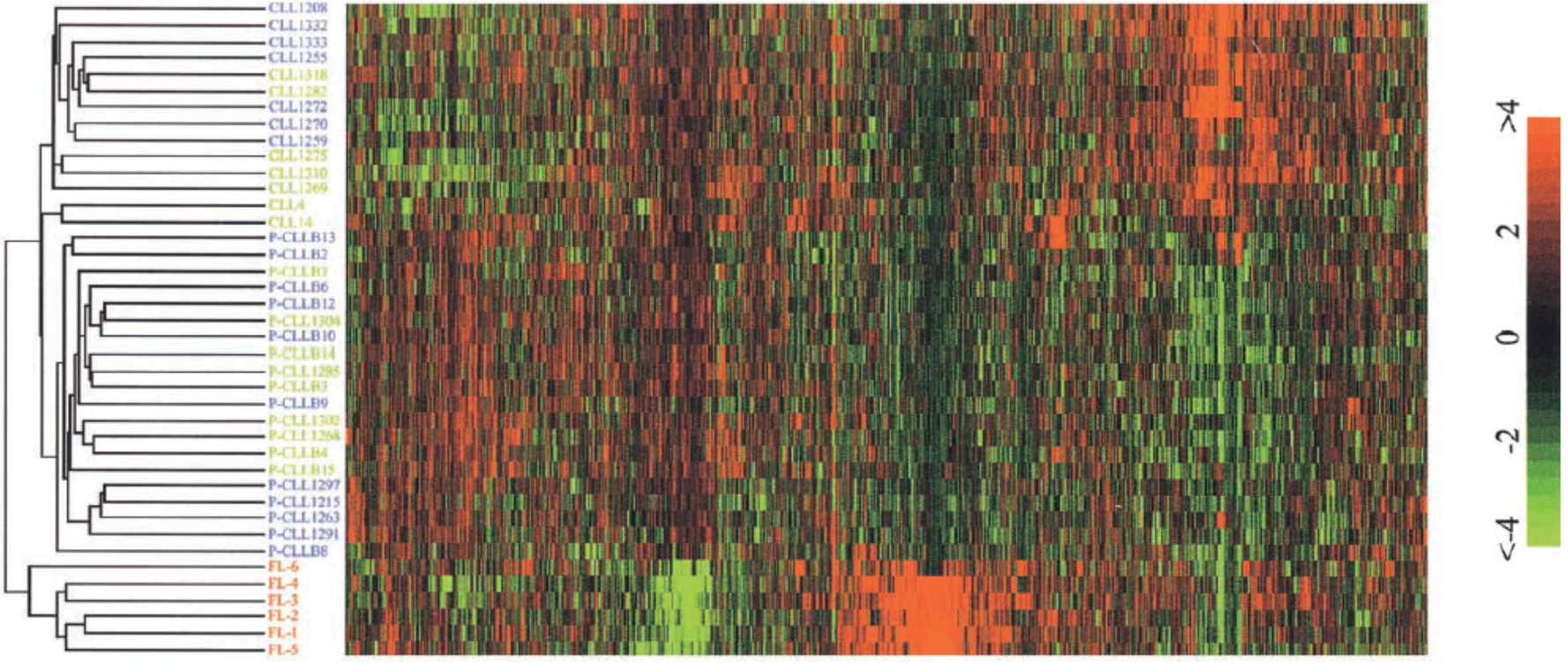
- 34 samples from patients with untreated CLL
  - 16 samples with unmutated IgV_H genes
  - 18 samples with mutated IgV_H genes
- 25 samples of normal B cells (NBC)
  - 5 samples, GC CD77$+$ from tonsil
  - 5 samples, GC CD77$-$ from tonsil
  - 5 samples, naive (pre-GC) from tonsil
  - 5 samples, memory (post-GC) from tonsil
  - 5 samples, GC-independent from umbilical cord blood
- 6 samples from patients with follicular lymphoma (FL)

Data processing

Note that the four kinds of B cells from tonsil are actually paired, since they were separated from the tonsils of only 5 individuals.

Arrays were quantified with MAS4.0 (average difference). They were truncated below at 20, then log transformed.

CLL can easily be distinguished from FL



Note, however, that the mutated (blue) and unmutated (green) CLL cases are intermingled.

Differential sample processing

By eye, one sees two prominent subclusters of CLL samples. These are distinguished in the names on the dendrogram by the prefix “P”.

- Twenty samples with the prefix “P” were purified by binding CD19-positive cells to magnetic beads. (11 mutated.)
- Fourteen samples without the prefix “P” came from patients in which the tumor cells accounted for more than 80% of the peripheral blood mononuclear cells, and were not purified.

Of course, that means we canot tell if differences between the two groups of samples are due to the severity of the disease (sicker patients being likely to have more tumor cells) or to the differential processing of the samples.

Further Notes on Clustering

Genes were filtered for large fold changes relative to the overall mean (2+ fold). This gave a subset of 2,337 “genes” (maybe probesets).

The paper says that clustering was performed using average linkage. To get distances, the profiles for each gene were first centered and scaled to have mean 0 and standard deviation 1, after which Euclidean distance was used.

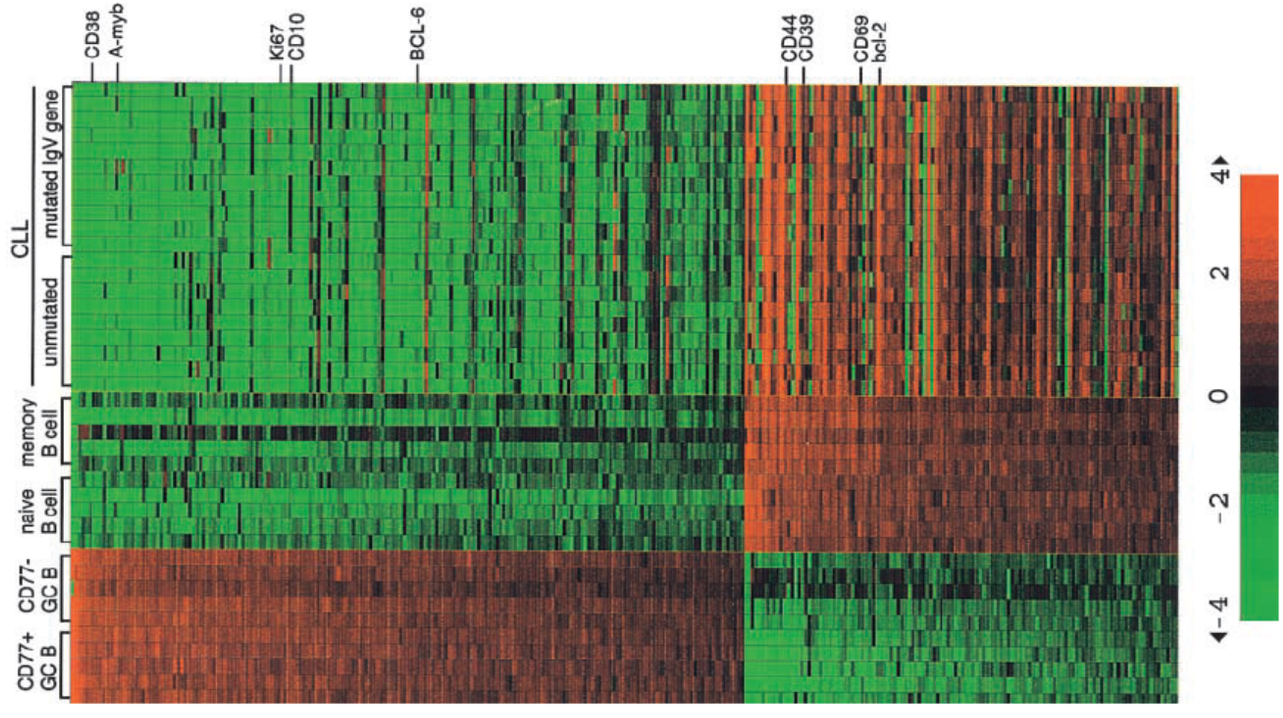
No statistics were used to test the robustness of the clusters.

Separating CLL subtypes

They use a modified t-statistic (difference in means divided by sum of standard deviations) to rank the genes for class comparison. Using this method on the on the 20 CLL cases where the samples were purified, they identified a set of 23 differentially expressed genes and used “weighted voting” to build a classifier. They applied this classifier to the 14 unpurified CLL samples and got correct answers on 12 out of 14 (the other 2 were ambiguous).

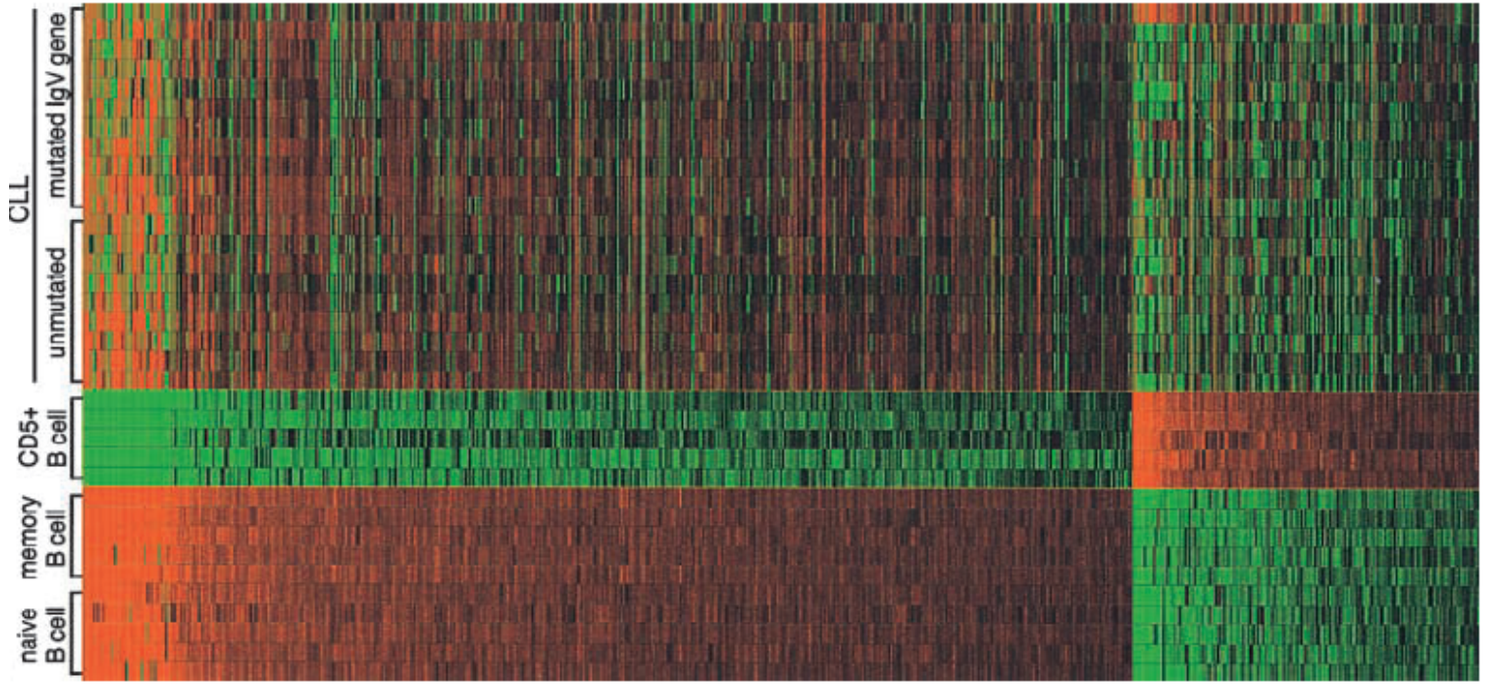
Of note, one of the 23 genes was “V4-31 Ig variable region”.

CLL does not look like germinal center B cells



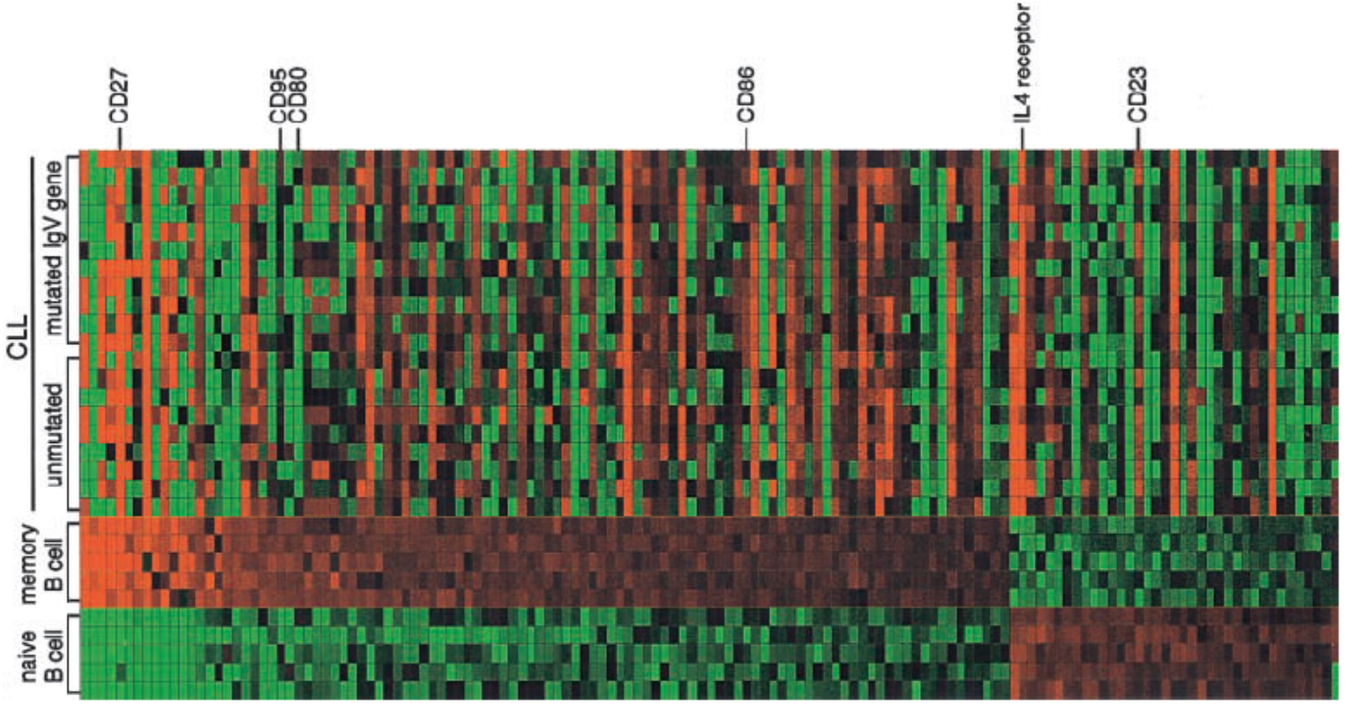
Genes were selected as differentially expressed between GC B-cells compared to naive or memory B cells. CLL displays a pattern similar to the non-GC B-cells. (Note: This analysis only uses the 20 purified CLL samples.)

CLL does not look like cord blood B cells



Genes were selected as differentially expressed between cord blood B-cells compared to naive or memory B cells. CLL displays a pattern similar to the non-GC B-cells. (Note: This analysis only uses the 20 purified CLL samples.)

CLL looks something like memory B cells



Genes were selected as differentially expressed between naive and memory B cells. The authors claim that 14 of the 20 purified CLL samples show a pattern similar to the memory B cells (6 others ambiguous). This similarity is independent of SHM.

Genes specific to CLL

Finally, the authors compared 10 randomly chosen CLL samples (5 mutated and 5 unmutated) to tonsillar B cells and to a collection of arrays on follicular lymphoma, Burkitt lymphoma, and diffuse large B-cell lymphoma and selected differentially expressed genes.

They did not explain why they only used a subset of the CLL samples.

The Lymphochip

Reference: Rosenwald A, et al. (2001) *Relation of gene expression phenotype to immunoglobulin mutation genotype in B cell chronic lymphocytic leukemia.* J Exp Med **194**: 1639–1647.

This study, from Lou Staudt's lab at the NCI in colaboration with Pat Brown at Stanford, also tried to understand the subtypes of CLL. They used glass arrays specifically designed for studying leukemia and lymphoma.

The Lymphochip: a custom microarray

The Lymphochip was first described in a well-known paper by Alizadeh et al. in Nature (2000; 403:503–511). They first built several libraries of cDNA clones from germinal center B cells, diffuse large B-cell lymphoma, follicular lymphoma, mantle cell lymphoma, and chronic lymphocytic leukemia. Clones were selected from these libraries and printed on glass arrays.

These arrays were used for two-color hybridizations with the experimental sample labeled with Cy5 and a common reference sample labeled with Cy3. The reference sample was made from a pool of RNA from nine different lymphoma cell lines.

A brief history of the Lymphochip

None of the papers describing results from the Lymphochip experiments explains that it has gone through multiple generations. These are not just re-printings of the same clones on a new set of glass slides. Instead, the researchers have repeatedly redesigned the array, and printed a different selection of clones in different places on the array.

You can recognize the different generations by the prefix on the file name (which presumably matches a barcode on the physical microarray). They have reported on data from at least five generations of the lymphochip (lc4b, lc5b, lc7b, lc8n, lc9n).

The “b” arrays only have 9,216 spots; the “n” arrays have 18,432 spots.

Samples used on the lymphochip

Rosenwald's study reports on 37 samples from untreated CLL patients. The mutation status was reported for only 28 samples (12 mutated, 16 unmutated). They also processed about 10 samples of normal B cells (NBC) obtained from various sites.

The CLL experiments were performed on the lc8n and lc9n arrays. All NBC experiments were performed on lc8n arrays; all CLL experiments were performed in lc9n arrays.

Data processing

Data was quantified using ScanAlyze (the usual tool for array studies coming out of Stanford). Assuming that processing was the same as in the Alizadeh paper, global normalization was applied to the log ratios to set the median equal to 1. Genes were filtered by intensity: they had to be 100 units above background in both channels or 500 units above background in at least one channel. Log ratios were used for the analysis.

Differential expression

They first compared CLL to a variety of arrays using normal B cells, normal T cells, DLBCL, and FL. Differential expression was assessed using two-sample t-statistics with unadjusted p-values ($p < 0.001$). They found 328 differentially expressed clones, corresponding to about 247 genes. Not surprisingly, these genes did not appear at all different in mutated vs. unmutated CLL samples.

Class prediction

Next, they randomly selected a training set to 10 unmutated and 8 mutated CLL samples. They again used two-sample t-statistics to select differentially expressed genes. They found 56 differentially expressed genes with unadjusted $p < 0.001$.

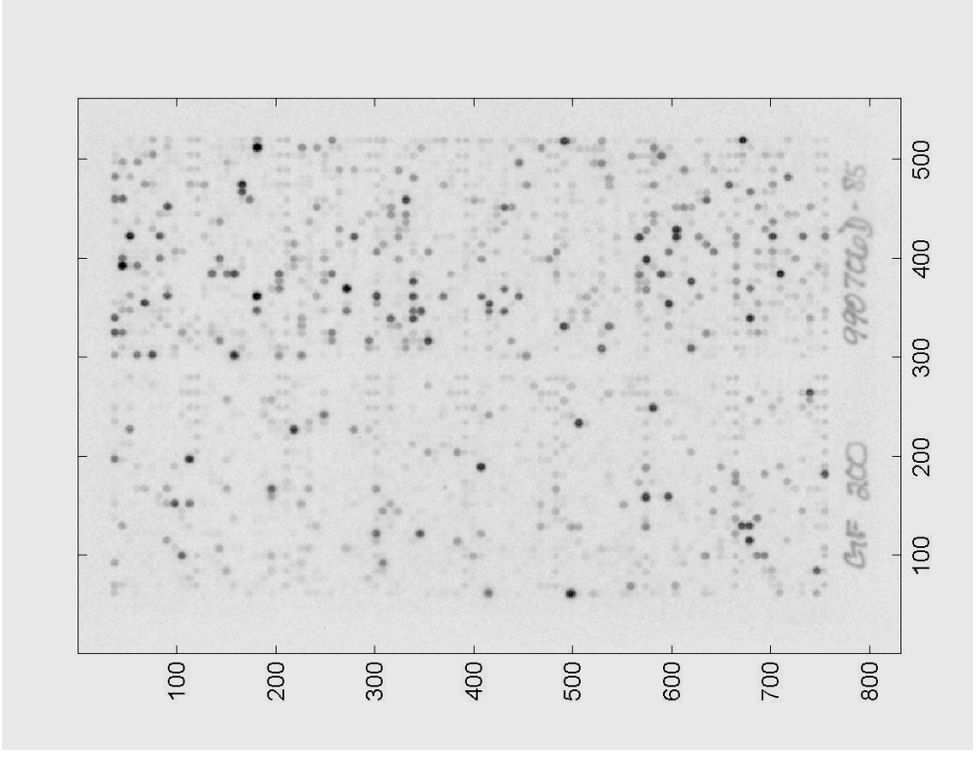
A class predictor was constructed using a linear combination of the log ratios for all 56 genes, weighted by the univariate t-statistics. The behavior of this procedure was tested using leave-one-out (applied to include the feature selection step). Leave-one-out accurately classified 17 out of 18 samples. The significance of the leave-one-out procedure was assessed by a permutation test. They got at least 17 out of 18 right in only 1 of 1000 permutations of the class labels. They also got the classification correct in 9 of the 10 CLL samples that were withheld from the training set. (ZAP70 alone is perfect.)

Class comparison

They next looked for differentially expressed genes between mutated and unmutated samples, and found 205 clones. Three samples were left out of this analysis (two because they had "low" levels of mutation, one because it just looked weird). Not surprisingly, hierarchical clustering using these 205 clones got everything except those three samples right.

Research Genetics GeneFilters

Early in the history of microarrays, several companies printed cDNA clones on nylon membranes. RNA samples were labeled with a radioactive isotope (typically ^{33}P), hybridized to the membrane, and scanned with a phosphorimager.



Samples used on GeneFilters

We ran six samples of untreated CLL (half mutated, half unmutated) and six samples of peripheral blood normal B cells from healthy adults. Each sample was hybridized to six different Research Genetics membranes (GF200 – GF205). Each type of membrane contains 5184 clones. In total, the six membranes have 31,104 cDNA clones representing 22,722 distinct UniGene clusters.

Data was quantified in ArrayVision, globally normalized by setting the 75^{th} percentile of the intensity to 1000, and log-transformed. A smoothed t-test (with Bonferroni-corrected p-values) was used to identify genes that were differentially expressed between CLL and NBC.

Meta-Analysis

Reference: Wang J, Coombes KR, Highsmith WE, Keating MJ, Abruzzo LV. (2004) *Differences in gene expression between B-cell chronic lymphocytic leukemiaand normal B cells: a meta-analysis of three microarray studies.* Bioinformatics, **20(17)**:3166-3178.

So, we took these three studies and asked ourselves how we could combine them. We focused on the problem of finding genes that were differentially expressed between CLL (of whatever subtype) and NBC (of whatever origin).

Quick question – how should we define a gene?

Data Processing

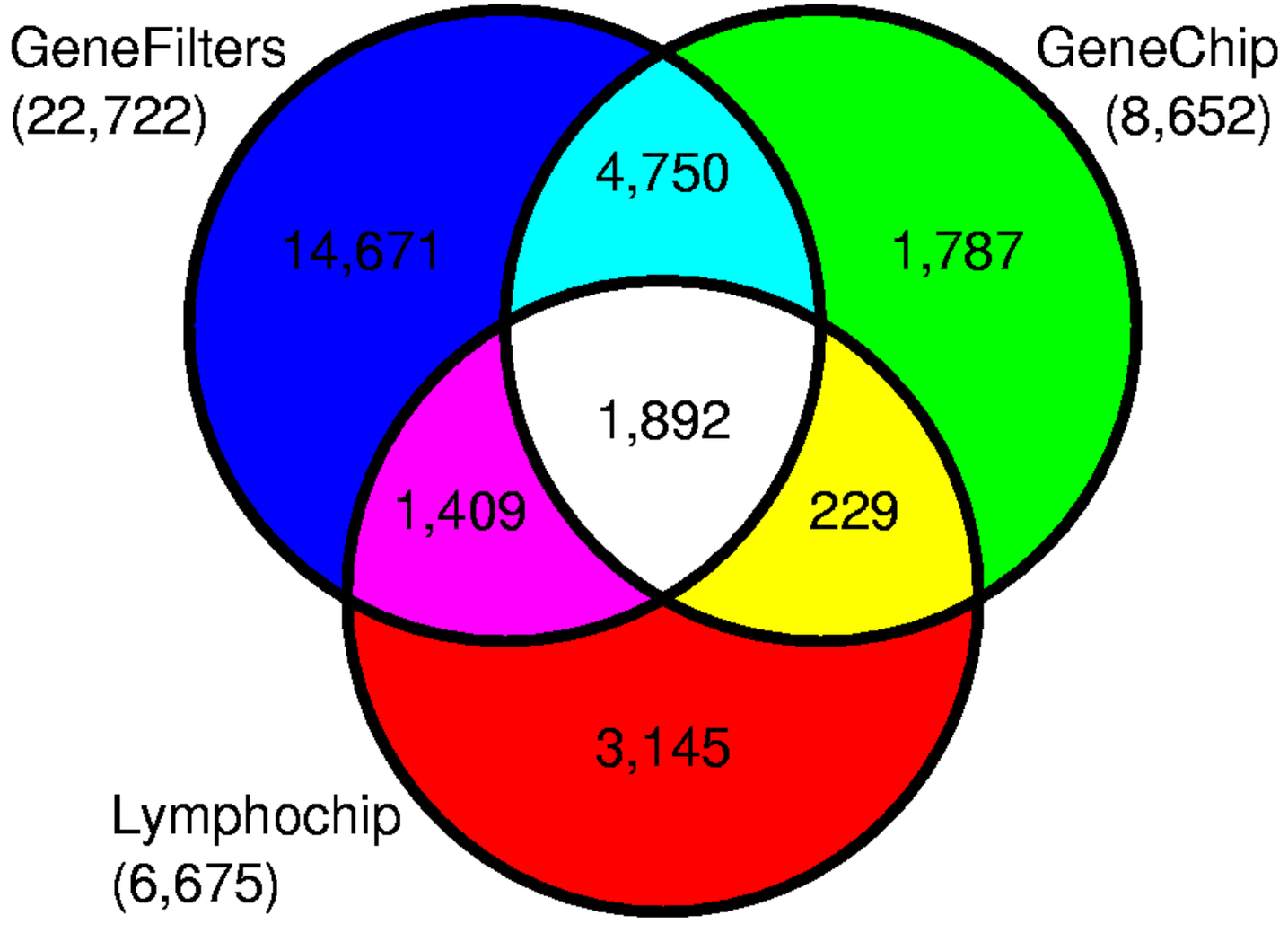
The Research Genetics data was processed as described above.

Since only the quantified Affymetrix data was made available (and not the CEL files), we used the data as provided.

We re-normalized the Lymphochip data, using a loess normalization and again scaling the 75^{th} percentile of the intensity to 1000. We also truncated all data below at 20, which was the value applied to the Affymetrix data before we received it. Log ratios of these values to the reference channel were then computed.

We also had to match spots on the two different generations of the Lymphochip. The lc8n and lc9n series had 15,497 clones in common, representing 6,675 distinct UniGene clusters.

Genes in common



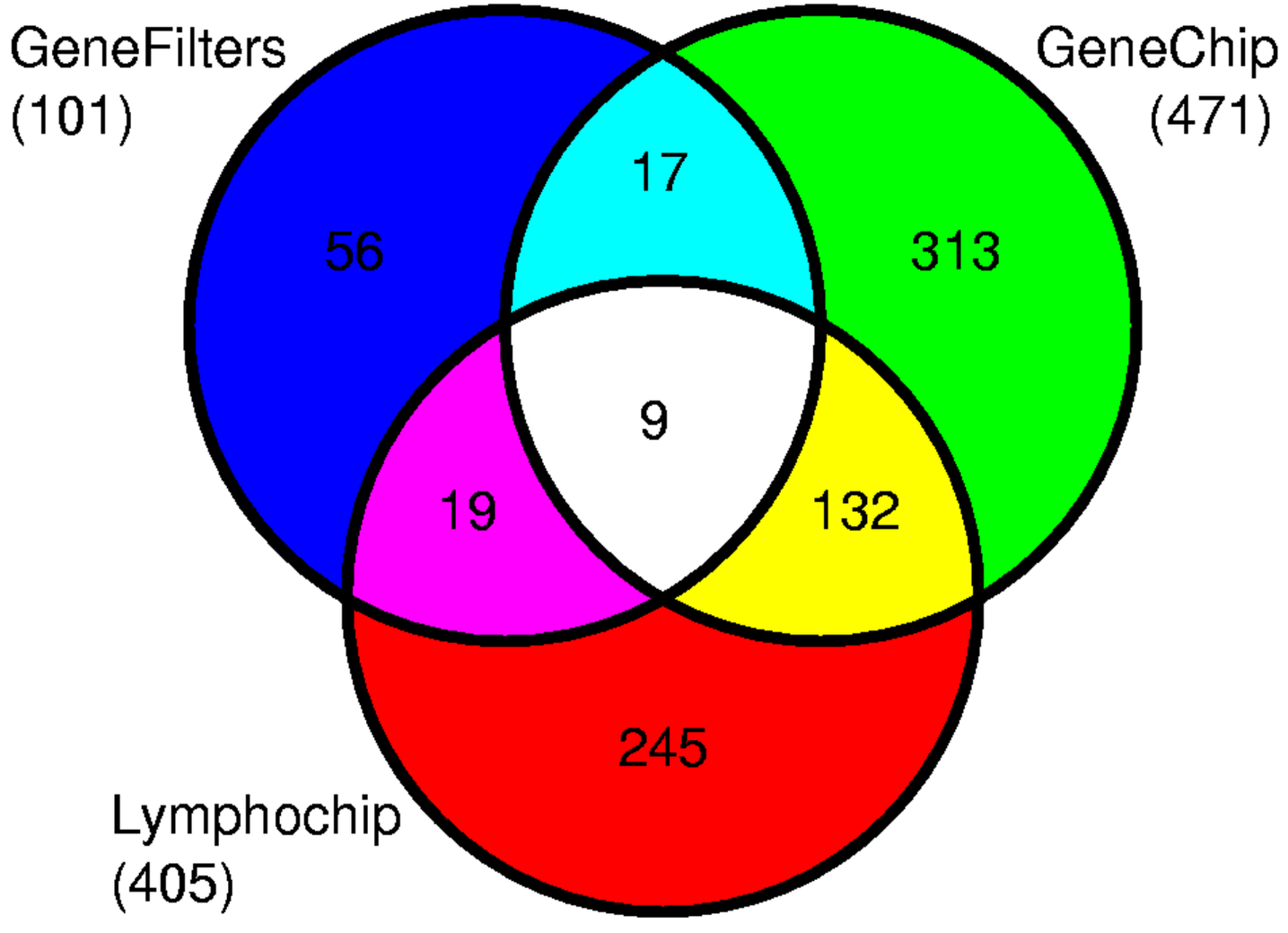
Subset of the samples

- Available data:
  - Affymetrix: 10 CLL, 20 NBC
  - Lymphochip: 33 CLL, 6 NBC
  - Research Genetics: 6 CLL, 6 NBC

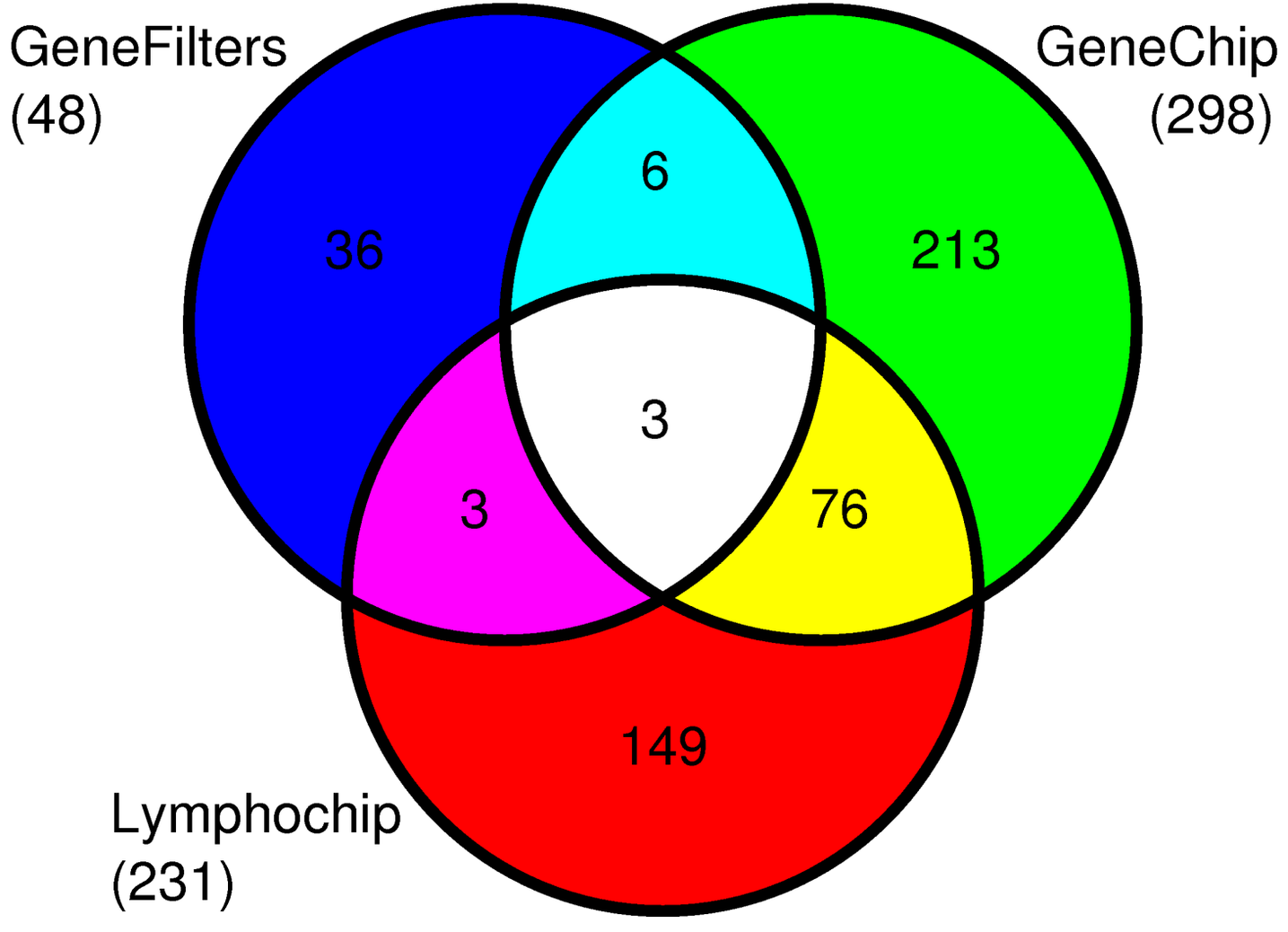
In order to get a balanced look at the platforms, we decided to use (a random selection of) 6 CLL and 6 NBC from each platform.

Differential expression was assessed on each platform separately using a smoothed t-test (i.e., the estimates of standard deviation are computed from a loess fit as a function of the mean log intensity), with a Bonferroni correction.

Comparing gene lists ($p < 0.001$)



Comparing gene lists ($p < 0.00001$)



Why is the agreement so bad?

Agreement between any two platforms is in the range of 25% – 30%, and that the agreement between all three is less than 10%.

Six samples versus six samples has poor power to detect true differences. Because the platforms use different probes with different affinities (and thus different variances), it is not terribly surprising that they do not find the same things.

There is also some difference in the normal B cell subsets. The Affymetrix and Lymphochip studies take NBCs from a variety of locations, including tonsil and cord blood. The Research Genetics study only used peripheral blood NBCs.

Can we do better?

Let's start by thinking about what happens on one platform. We have (for each gene) independent measurements of X_C (the log expression in CLL) and X_B (the log expression in NBC). We assume that

$$X_C \sim N(\mu_C, \sigma_C), \qquad X_B \sim N(\mu_B, \sigma_B).$$

We are interested in estimating the parameter

$$\delta = \mu_C - \mu_B,$$

which is the logarithmic fold-change in expression. The natural estimate, of course, is just

$$D = \bar{X}_C - \bar{X}_B.$$

The test statistic

Now, to keep things general, suppose we observe log expression values from n_C CLL samples and n_B NBC samples. Then the unknown parameter δ is normally distributed with mean D and variance determined by

$$\sigma_D^2 = \frac{\sigma_C^2}{n_C} + \frac{\sigma_B^2}{n_B}.$$

To perform a hypothesis test on the difference of means when the standard deviation is known, the appropriate test statistic is

$$Z = \frac{D}{\sigma_D} = \frac{\bar{X}_C - \bar{X}_B}{\sqrt{\frac{\sigma_C^2}{n_C} + \frac{\sigma_B^2}{n_B}}}.$$

Two key observations

1. When using the smooth t-test, the standard deviation is estimated using a huge number of data points. For all practical purposes, we can treat the σ's obtained this way as "known".

2. All microarray platforms yield estimates of δ; they do so with different precision.

Combining measurements with different precision

There is a standard way to combine measurements of the same quantity made with instruments of different precision: you weight each estimate by its variance.

For instance, let $D_L = \bar{X}_{C,L} - \bar{X}_{B,L}$ be the estimate of δ on the Lymphochip platform, with variance σ_L^2 computed as above. Let D_A and σ_A be the corresponding quantities on the Affymetrix platform. Then the combined estimate is

$$D_{combined} = \frac{(D_L/\sigma_L^2) + (D_A/\sigma_A^2)}{(1/\sigma_L^2) + (1/\sigma_A^2)}$$

with variance computed from

$$\frac{1}{\sigma_{combined}^2} = \frac{1}{\sigma_L^2} + \frac{1}{\sigma_A^2}.$$

Combining measurements with different precision

This formula generalizes immediately to more than two platforms. If D_R and σ_R are the corresponding quantities from the Research Genetics microarrays, then

$$D_{combined} = \frac{(D_L/\sigma_L^2) + (D_A/\sigma_A^2) + (D_R/\sigma_R^2)}{(1/\sigma_L^2) + (1/\sigma_A^2) + (1/\sigma_R^2)}$$

with variance computed from

$$\frac{1}{\sigma_{combined}^2} = \frac{1}{\sigma_L^2} + \frac{1}{\sigma_A^2} + \frac{1}{\sigma_R^2}.$$

The correct test statistic on the combined data is

$$Z_{combined} = \frac{D_{combined}}{\sigma_{combined}} = \frac{D_L}{\sigma_L^2} + \frac{D_A}{\sigma_A^2} + \frac{D_R}{\sigma_R^2}$$

Remarks

These formulas do not depend on having equal numbers of samples on each platform; they automatically adjust for the number of samples used.

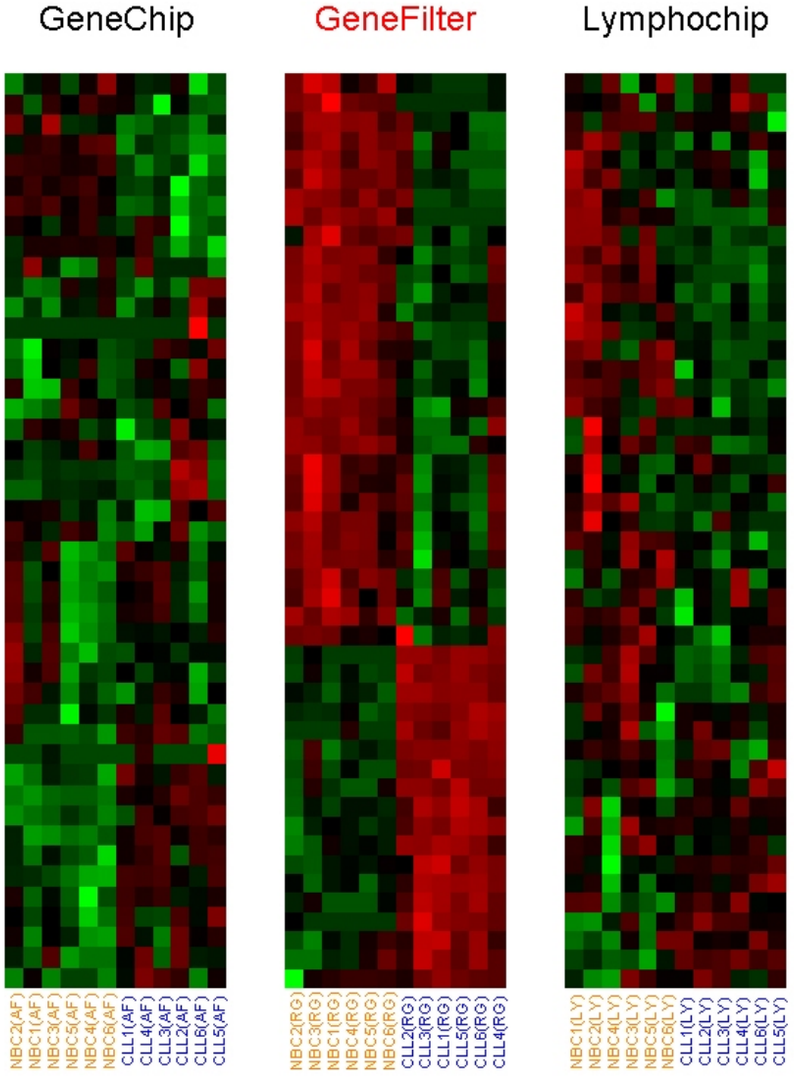
Also note that the final formula does not depend on the order in which the platforms were combined.

How well does this work?

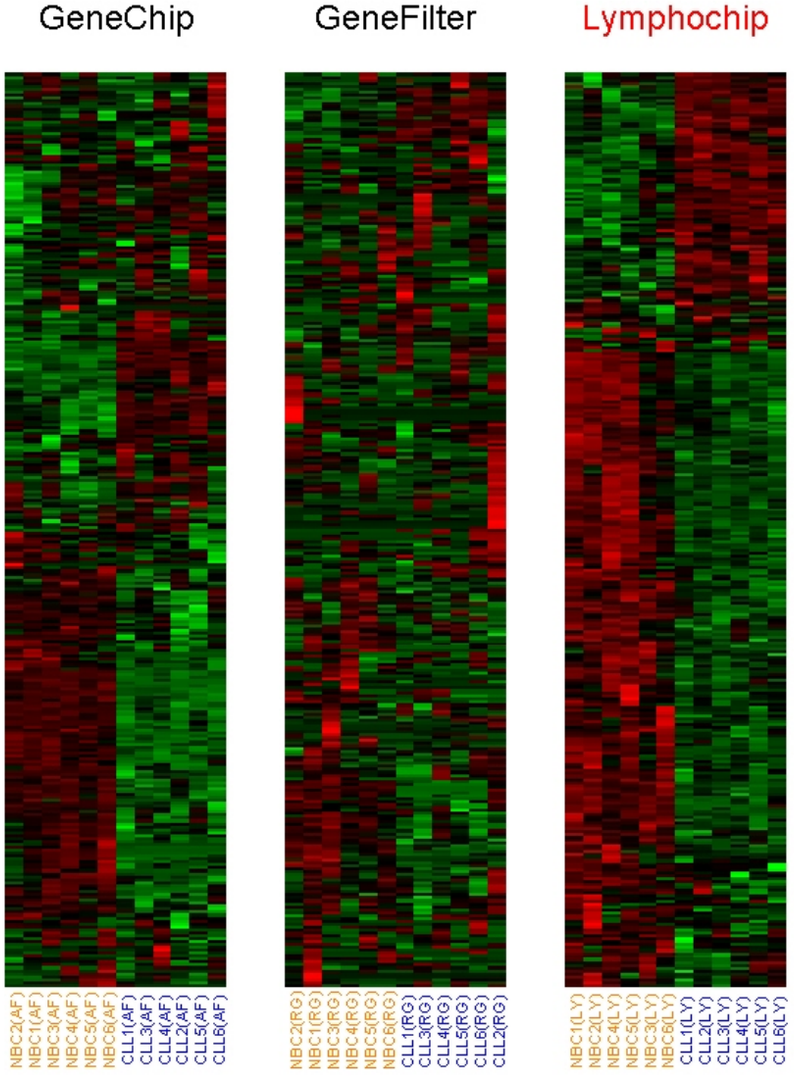
When we used this method to combine the data from CLL and NBC samples on the three platforms, we found that 124 genes were differentially expressed (setting an extrmely high cutoff: $|Z| > 8$). In a PubMed search, we found that 20 of these 124 genes had previously been reported to be differentially expressed between CLL and NBC using other technologies, and that 19 of the 20 genes changed expression in the direction compatible with the literature.

We also looked at the functional categories of the genes identified as different. These were significantly enriched for genes involved in “response to external stimulus, “stress response”, and “apoptosis”, all of which make sense for CLL.

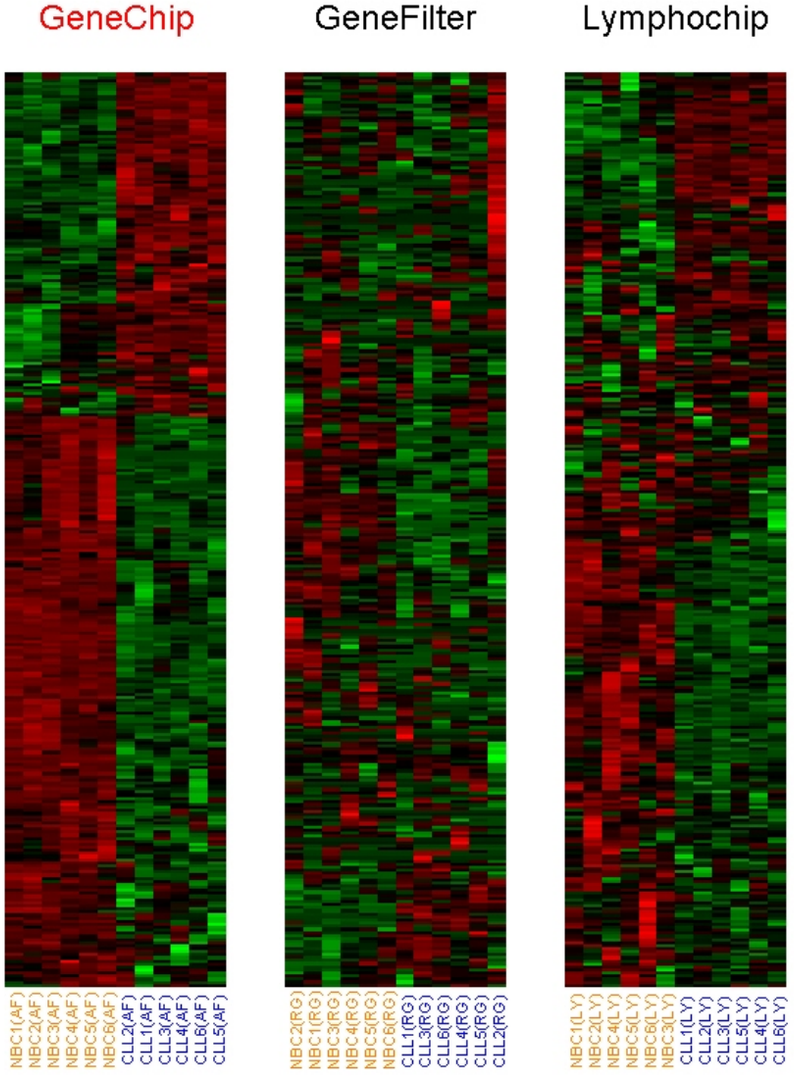
Two-way clustering: Research Genetics



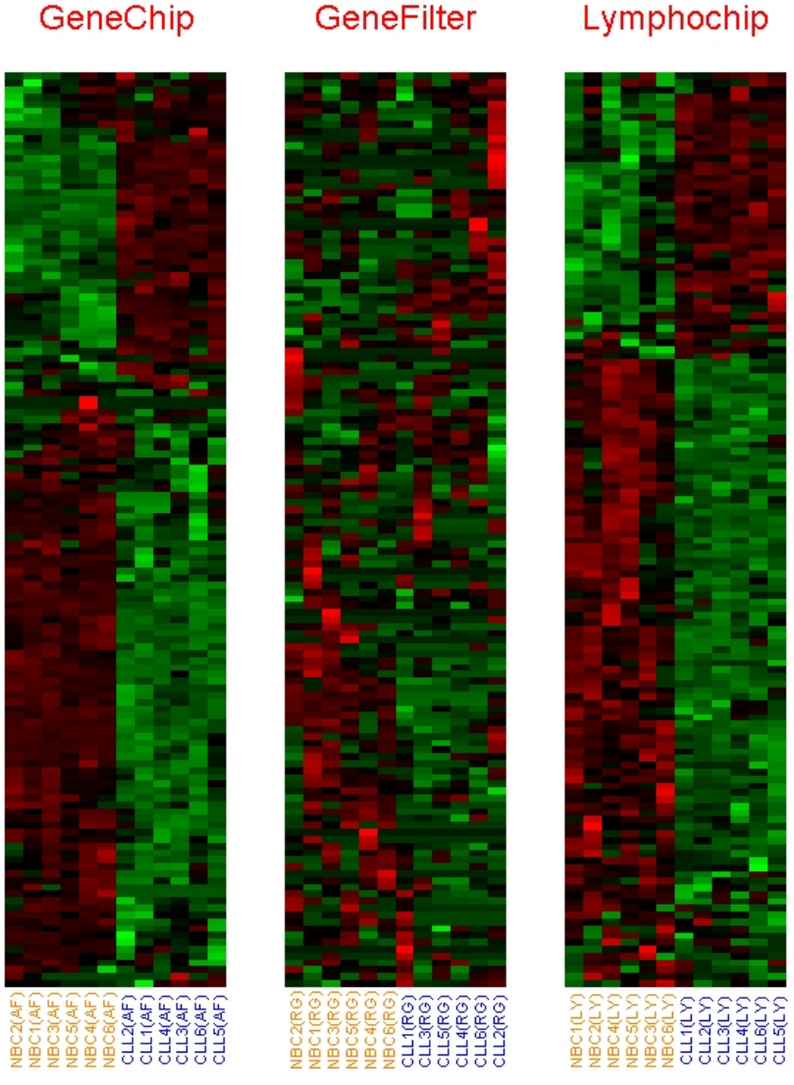
Two-way clustering: Lymphochip



Two-way clustering: Affymetrix



Two-way clustering: Meta-Analysis



Predicting Mutation Status

- Goal: Find a way to predict prognosis of CLL patients without having to sequence IGVH genes
- Idea: Settle for predicting SHM status, as defined by directing sequencing, as a “gold standard”.
- Idea: Select potential predictors from multiple microarray studies. Used papers by Klein and by Rosenwald. Also used lymphochip arrays from Wiestner et al., *Blood*, 2003 and U133A arrays from Abruzzo et al., *J Mol Diagn*, 2005.

Reference: Abruzzo et al., Identification and validation of biomarkers of IgVH mutation status in chronic lymphocytic leukemia using microfluidics QRT-PCR technology. *J Mol Diagn*. 2007; 9:546-55

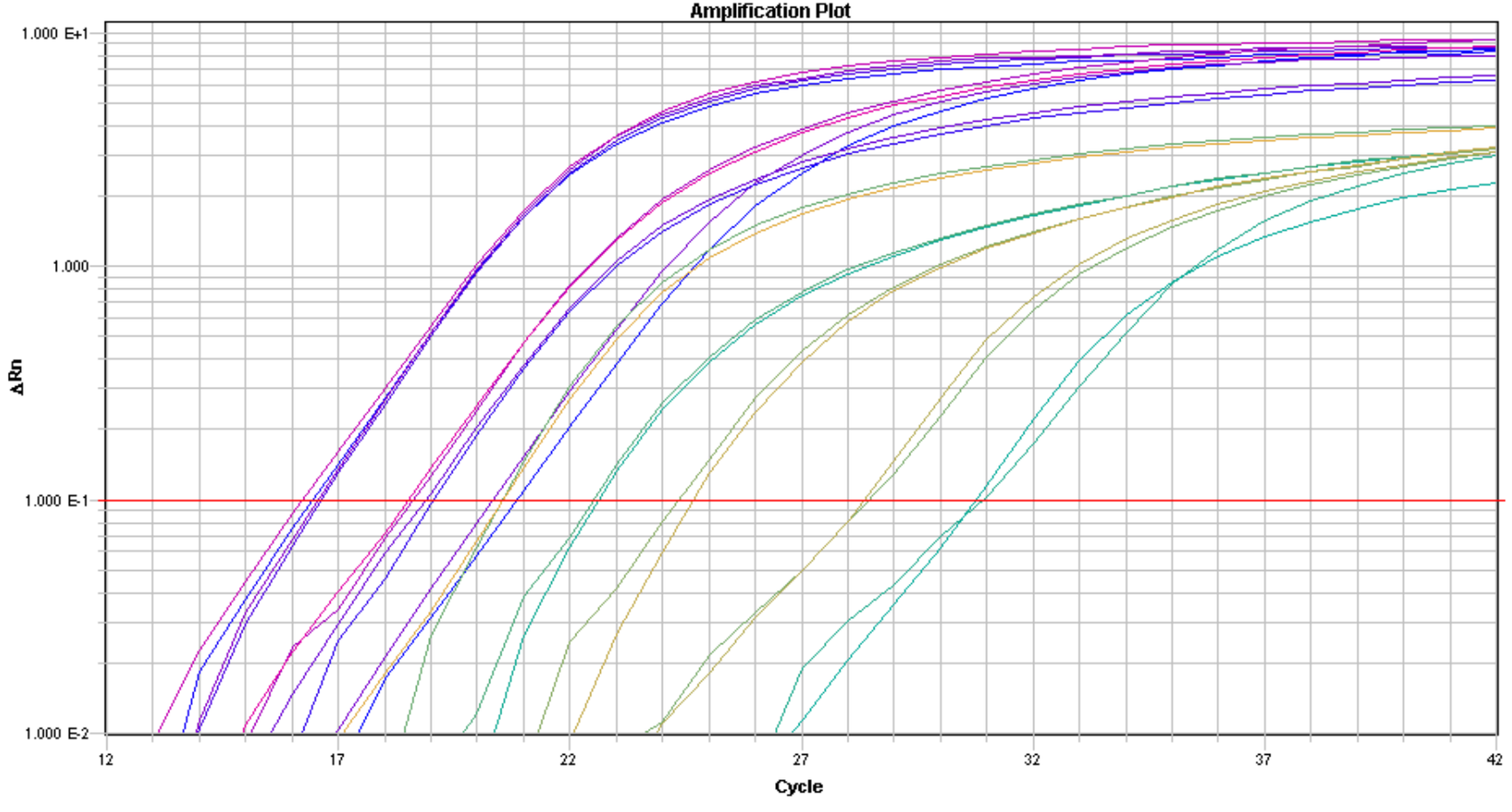
Selecting candidate predictors

We analyzed the studies separately and jointly, using tools you have already seen: t-tests, Wilcoxon rank sum tests, BUM, epirical Bayes, and the above meta-analysis.

We kept any gene that showed up in more than one study. We kept genes that showed up in only one study if the evidence appeared particularly strong. We kept genes found in the meta-analysis, and produced a list of 88 differentially expressed genes.

Real-time PCR data

The selected candidates were measured (in parallel) using real-time PCR.



Real-time PCR data

For each gene and each sample, the real-time PCR measurement is quantified by recording the number of cycles C_T at which the opbserved fluorescence reaches a fixed threshold.

Data are normalized by subtracting the mean C_T value for a set of housekeeping genes: `PGK1`, `18s rRNA`, `GUSB`, `ECE1`, and `GAPDH`. (These genes were selected using a preliminary experiment that identified them as having the most constant expression across a set of CLL samples. See Abruzzo et al., *Biotechniques*, 2005.)

Experimental Design: Training Set

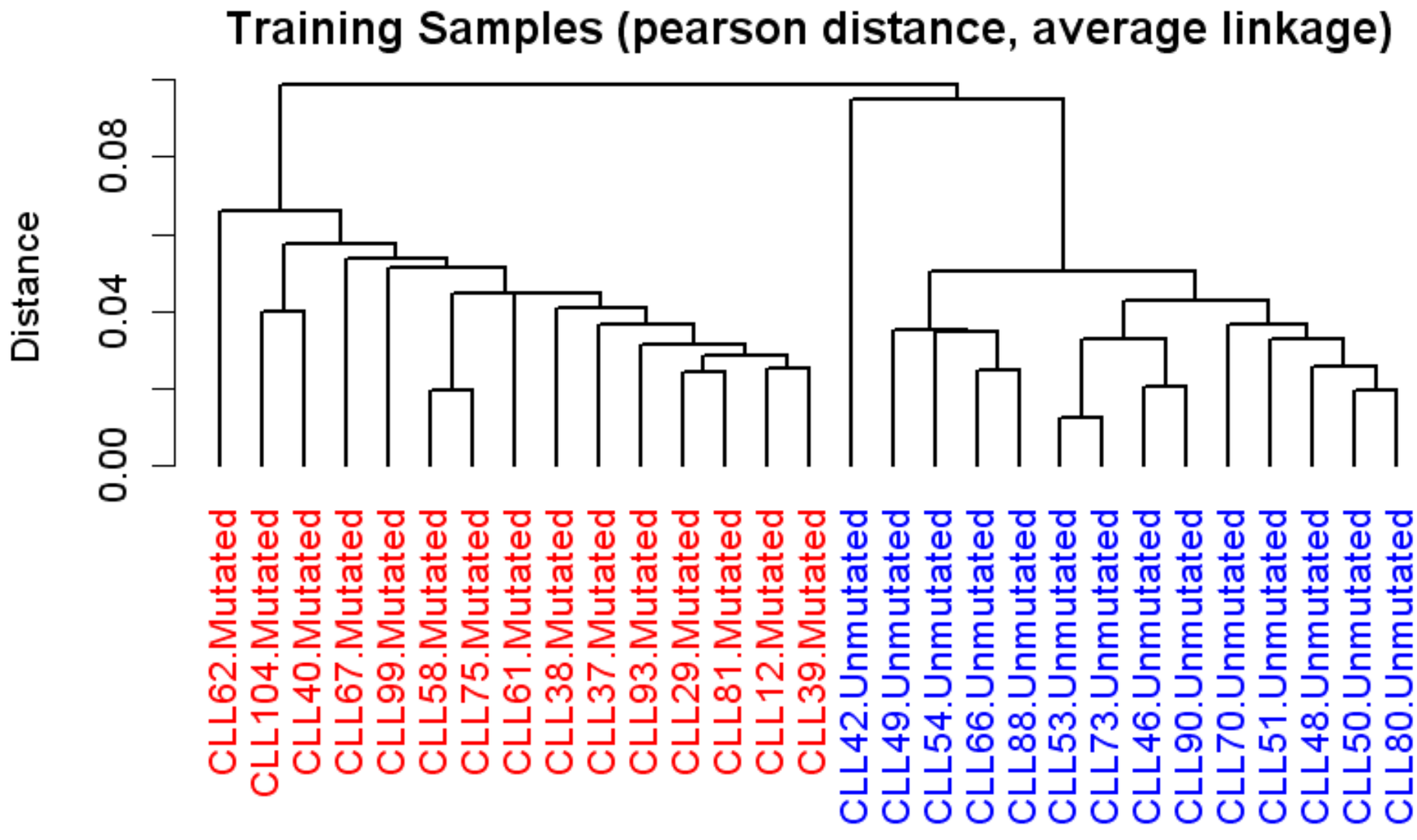
- Select 15 mutated and 15 unmutated CLL samples.
- Run each one on a microfluidics card to measure 94 genes (88 candidates and 8 controls) in duplicate.
- Average the replicates.
- Normalize to the mean of 5 housekeeping genes.
- Should get a 96×30 matrix.

Experimental Results: Training Set

Things do not always work out as you plan:

- One sample failed QC
- Primer sets for two genes produced no data
- Actually ended up with a 94×29 matrix .

Experimental Results: Training Set



Experimental Results: Training Set

Using the 29 samples in the training data, we also performed univariate t-tests to see how many of the 86 candidate genes actually appeared to be differentially expressed based on the real-time PCR data on this new set of samples. We confirmed the differential expression of $37/86 = 43\%$ of these genes.

Note that we later repeated this analysis using 49 samples from a combined training and test set, and confirmed the differntial expression of a total of $48/88 = 56\%$.

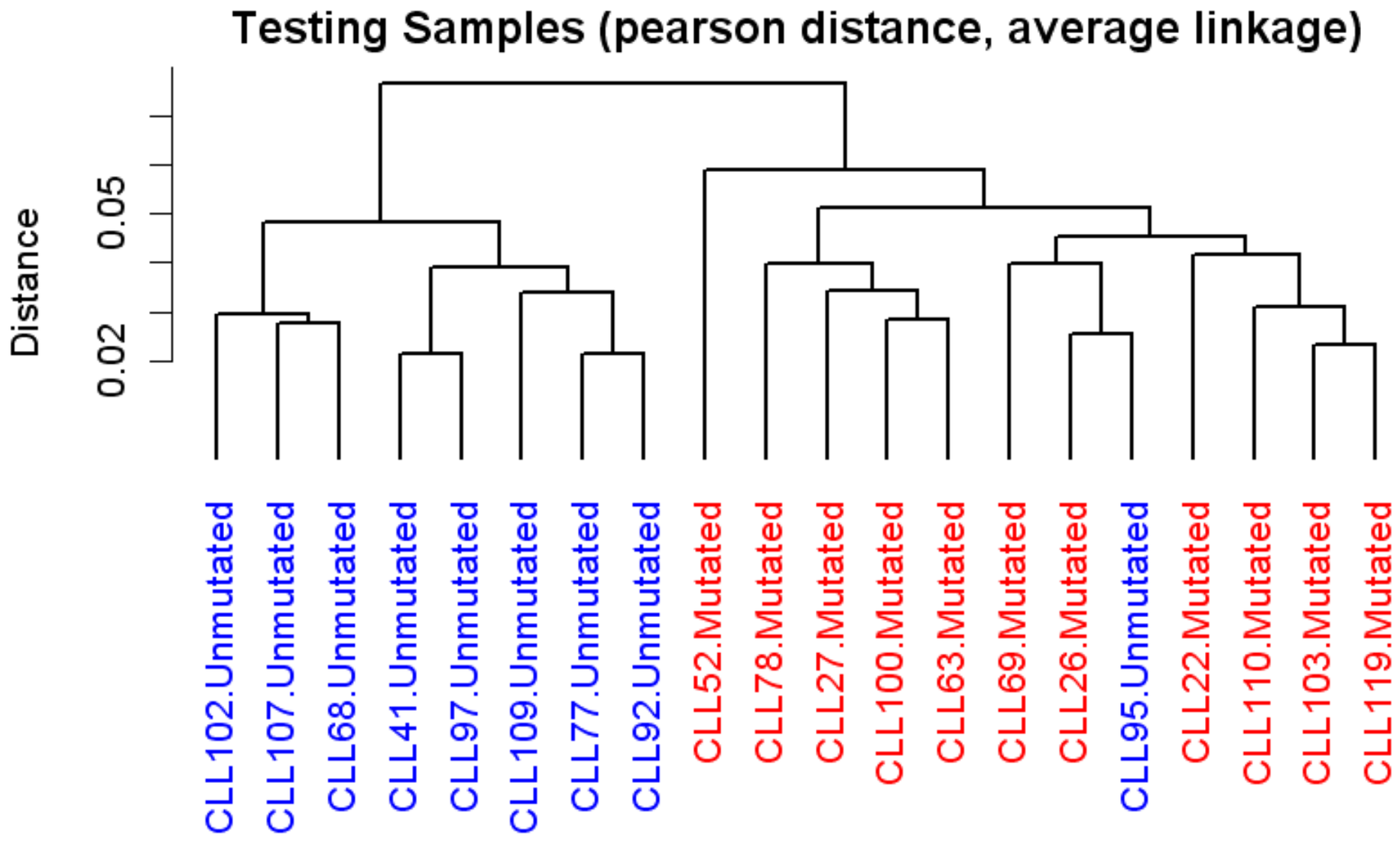
Even using multiple microarray studies and correcting for multiple testing, only about half of the candidates could be confirmed as differentially expressed on a new data set using more accurate technology.

Experimental Design: Training Models

Using the training set, we used 16 different methods to construct models that could predicxt the mutation status based on the mRNA profiles measured using real-time PCR. All models were completely conmstructed before the test samples were run on the microfluidics cards.

We then ran 20 test samples. Data was processed the same way as before.

Experimental Results: Test Set

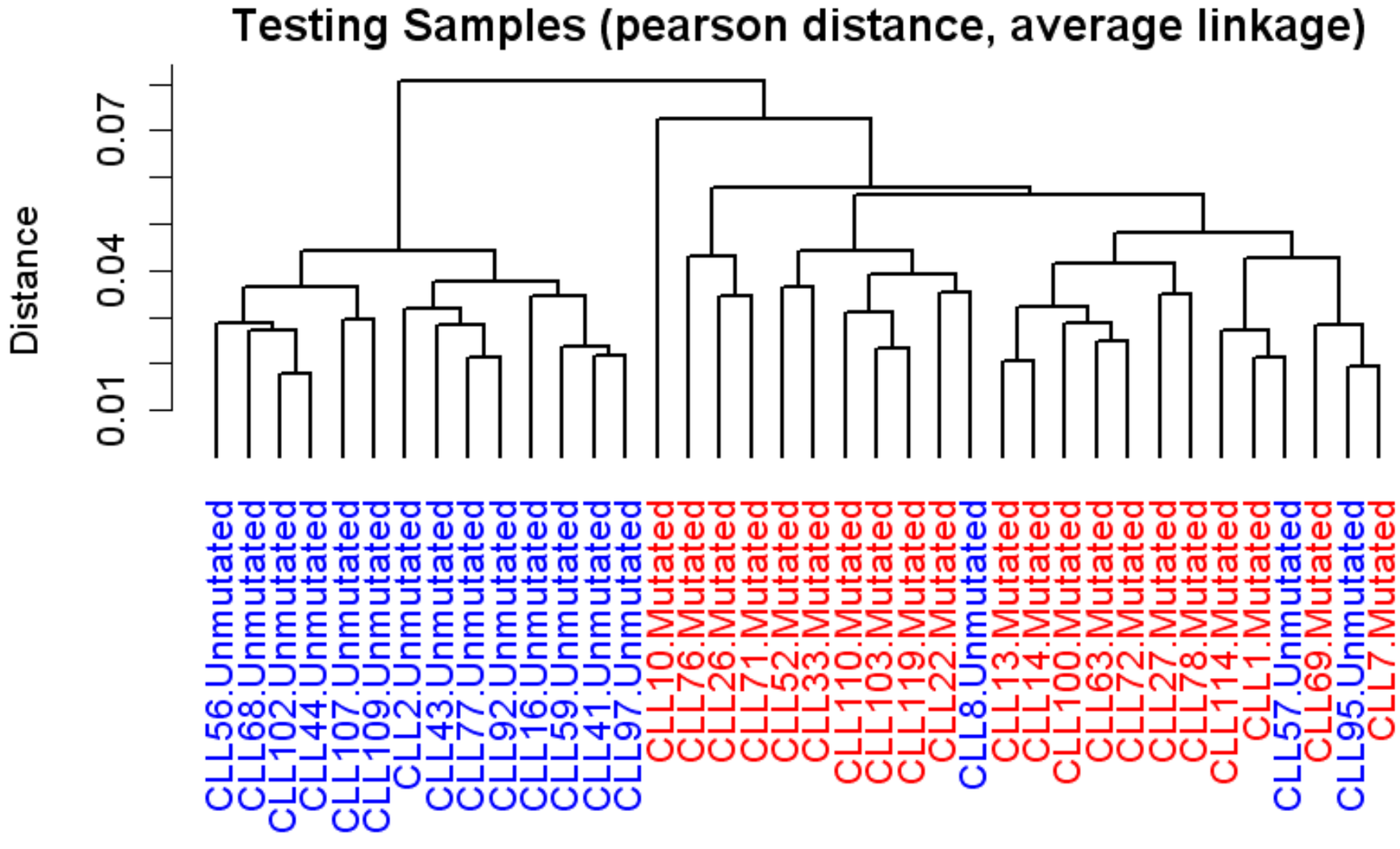


Experimental Results: Testing Predictions

ID	Classifier	Feature Selection	Train Mut	Train Unmut	Test Mut	Test Unmut
1	QDA	Top 3, t-test	15/15	14/14	11/11	8/9
2	LDA	Top 4, t-test	15/15	14/14	11/11	8/9
3	Mol Signs (3/7)	Top 7, tail-rank	15/15	14/14	11/11	8/9
4	Mol Signs (2/4)	Top 4, tail-rank	15/15	14/14	10/11	8/9
5	DLDA	Wrapper (24)	15/15	14/14	11/11	8/9
6	CCP	All genes	15/15	14/14	11/11	8/9
7	CCP	Top 4, t-test	15/15	14/14	11/11	7/9
8	KNN (k=3)	All genes	15/15	14/14	11/11	8/9
9	KNN (k=3)	Top 4, t-test	15/15	14/14	11/11	6/9
10	NN Ensemble	All genes	15/15	14/14	11/11	8/9
11	LDA	Wrapper (3)	15/15	14/14	7/11	8/9
12	Naive Bayes	All genes	15/15	14/14	11/11	8/9
13	PCR (k=1)	All genes	15/15	14/14	9/11	8/9
14	Random Forest	All genes	15/15	14/14	11/11	8/9
15	CART	Wrapper(2/94)	15/15	14/14	10/11	7/9
16	CART	Wrapper(3/19)	15/15	14/14	10/11	7/9

More Tests

Since the paper was published, we have run 18 more samples.



More Tests

ID	Classifier	Feature Selection	Train Mut	Train Unmut	Test Mut	Test Unmut
1	QDA	Top 3, t-test	15/15	14/14	21/21	14/17
2	LDA	Top 4, t-test	15/15	14/14	21/21	14/17
3	Mol Signs (3/7)	Top 7, tail-rank	15/15	14/14	21/21	14/17
4	Mol Signs (2/4)	Top 4, tail-rank	15/15	14/14	19/21	14/17
5	DLDA	Wrapper (24/94)	15/15	14/14	21/21	14/17
6	CCP	All genes	15/15	14/14	21/21	15/17
7	CCP	Top 4, t-test	15/15	14/14	21/21	13/17
8	KNN (k=3)	All genes	15/15	14/14	21/21	14/17
9	KNN (k=3)	Top 4, t-test	15/15	14/14	21/21	12/17
10	NN Ensemble	All genes	15/15	14/14	21/21	14/17
21	LDA	Wrapper (3/94)	15/15	14/14	17/21	15/17
12	Naive Bayes	All genes	15/15	14/14	21/21	15/17
13	PCR (k=1)	All genes	15/15	14/14	19/21	14/17
14	Random Forest	All genes	15/15	14/14	21/21	14/17
15	CART	Wrapper(2/94)	15/15	14/14	20/21	13/17
16	CART	Wrapper(3/19)	15/15	14/14	20/21	13/17

Multiple Testing

We built 16 different models on the trainin data, and we have applied all of them to the test data. Should there be some kind of penalty for "multiple testing" because of all these models?

Note, however, that all models were constructed before the first test set was analyzed, and long before the second test set was collected.

All models find some structure in the data that is better than chance. (As does the unsupervised clustering.)

Worst performance: $32/38 = 84.2\%$ accuracy.

Best performance: $36/38 = 94.7\%$ accuracy.

Median performance: $35/38 = 92.1\%$ accuracy.

Conclusions

1. Feature selection on microarrays is hard. We only confirmed slightly more than half of the genes that were supposed to be differentially expressed.

2. Feature selection on microarrays works. The dominant signal in the PCR data was the difference between mutated and unmutated samples.

3. With good features available, classification is easy. Many different classification methods worked about equally well on the PCR data.