Biology 644

Old Title: Bioinformatics for Molecular Biologists

Potential New Title: Integrated Bioinformatics Using R for Both Wet and Dry Scientists

Chromosomal Abnormalities

46,XX - Normal Female Karyotype
46,XY - Normal Male Karyotype

- These descriptions say there are 46 chromosomes and that it is a male or female.

46,XX,del(14)(q23)

- Female with 46 chromosomes with a deletion of chromosome 14 on the long arm (q) at band 23. The numbers after p or q refers to regions, bands and subbands seen when staining the chromosome with a staining dye

46,XY,dup(14)(q22q25)

- Male with 46 chromosomes with a duplication of chromosome 14 on the long arm (q) involving bands 22 to 25.

46,XX,r(7)(p22q36)

- Female with 46 chromosomes with a 7 chromosome ring. The end of the short arm (p22) has fused to the end of the long arm (q36) forming a circle or 'ring'

47,XY,+21

- Male with 47 instead of 46 chromosomes and the extra chromosome is a 21. (Down Syndrome)

46,XX,t(9;14)(p1q23)

- Female with 46 chromosomes with a translocation between chromosomes 9 and 14; short arm (p) band 1 and long arm (q) band 23.

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

- Training Set = 33 adult patients with T-cell acute lymphocytic leukemia (T-ALL)
  - 6 were refractory to induction chemotherapy
  - 2 died during induction chemotherapy,
  - 25 achieved complete remission (CR)
- Test Set = 18 adult patients
- RNA prepared from bone marrow and peripheral blood mononuclear cells
- Affymetrix HGU95aV2 microarray containing probes for ~10K genes
- Unsupervised Learning to discover T cell differentiation
  - hierarchical clustering of the 33 T-ALL samples based on expression of 313 selected
- Supervised Learning to classify responding versus refractory
  - Classifier containing 34 genes
- Supervised Learning to classify relapsed versus continuous complete remission
  - Classifier containing 19 genes
  - Classifier containing just 3 genes

Table 1. Characteristics of T-ALL patients

	Gene expression profile set (n = 33)	RT quantitative PCR set (n = 18)
Clinical characteristics		
Male/Female	24/9	14/4
Age, y (range)	29.5 (14-52)	22.5 (14-37)
WBC × 10^9/L (range)	87 (3-700)	52 (6.8-848)
Immunophenotypic characteristics		
T1	2	0
T2	15	7
T3	10	7
T4	2	1
Incomplete phenotype	4	3
Cytogenetic characteristics		
Normal karyotype	8	8
Not evaluable	13	8
t(4;11)*	1	0
t(10;14)	1	0
del(6q)	2	1
del(7q)	1	0
Other simple alterations:	4†	1‡
Other complex alterations:	3§	0

*t(4;11)(q21;p15). Molecular studies detected the presence of NUP/RAP molecular rearrangement.
†dup(12) (q12;q24); add(19)(p13); del(11q)(q21); del(4)(p14).
‡del(3), del(17).
§del(2)(p16), add(7)(p22), +22; del(7)(p15), t(9;12)(q12;q32), del(13)(q13;q33); der(1;13)(q10;p12),+1,+4,+18+21,+22.

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

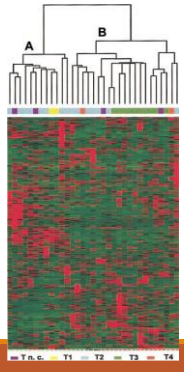


Figure 1. Unsupervised hierarchical clustering of the 33 T-ALL samples based on expression of 313 selected genes.

- **Each column represents a sample and each row represent a gene.**
- **Relative levels of gene expression are depicted with a color scale where red represents the highest level of expression and green represents the lowest level.**
- **Unsupervised clustering identified 2 subsets of samples, A and B.**
- **A Fisher's exact test for association between clusters (group A and B) and levels of T-cell differentiation was performed and proved highly significant (P.01).**

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

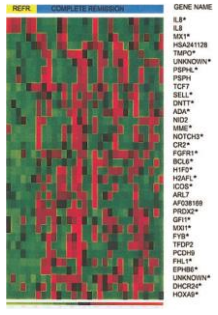


Figure 2. Expression of the top 34 selected genes in responding versus refractory T-ALL patients (supervised learning).

- Expression of 34 selected genes in T-ALL from 6 patients who did not respond to induction therapy and 25 patients who achieved CR.
- To select genes that were differentially expressed in subgroups of interest, the t test was applied and genes selected based on the nominal P values attained. (a t test with P cut-off of .05 was the criterion used to select a gene as best performing)
- Asterisks identify the top 25 genes that provide the smallest prediction error.

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

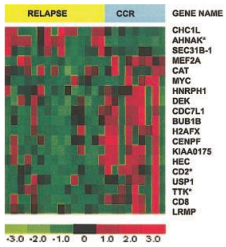


Figure 3. Expression of 19 selected genes in T-ALL patients who relapsed versus those who remained in CCR.

- Expression of 19 elected genes in T-ALL from 8 patients who remained in CCR for more than 2 years and 16 patients who experienced a relapse less than 2 years after achieving CR.
- Asterisks identify the top 3 genes that provide the smallest prediction error.

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

Table 2. Mean gene expression values for the 3 best predictor genes, *AHNAK*, *CD2*, and *TTK*

	AHNAK	*CD2*	*TTK*
CCR	747	478	195
Relapse	1954	261	126

Normalized expression values from Affymetrix arrays.

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

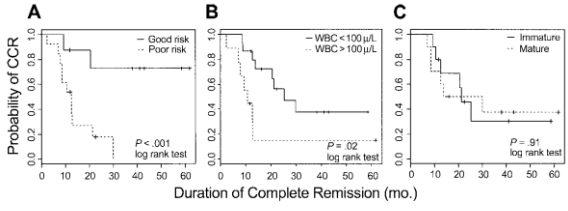
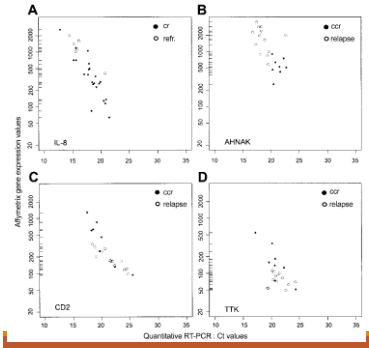


Figure 4. Kaplan-Meier plots estimating probability of maintaining CR for adult T-ALL.

(A) 24 evaluable patients were assigned to either good-risk or poor risk T-ALL based on expression of *AHNAK*, *CD2*, and *TTK* as measured by oligonucleotide microarrays.
(B) Kaplan-Meier plots based on the WBC count at diagnosis.
(C) Kaplan-Meier plots based on the degree of T-lineage differentiation of the leukemic cell (immature T1-T2; mature T3-T4 of EGIL classification).

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004



- **Figure 5. Gene expression values in adult T-ALL measured by oligonucleotide microarray and quantitative RT-PCR.**
- For ***IL8***, ○ represents refractory patients, and ● represents patients who achieved CR.
- For ***AHNAK***, ***CD2***, **and** ***TTK***, ○ represents patients who experienced a relapse, and ● represents patients who remain in CCR.

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

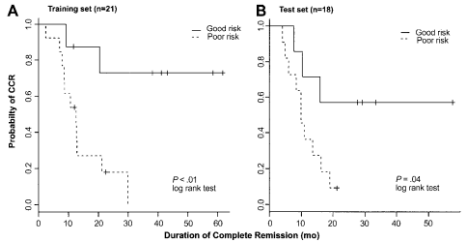


Figure 6. Kaplan-Meier plots. Kaplan-Meier plots represent probability of maintaining CR in a **training set of 21 patients (A)** and a **test set of 18 patients (B)** with T-ALL treated on the same clinical protocol. Patients were assigned to either good-risk or poor-risk T-ALL based on expression of ***AHNAK***, ***CD2***, **and** ***TTK*** as measured by **RT-PCR**.

"Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival" – Chiaretti et al., Blood 2004

- **Linear Discriminant Analysis (LDA)** - closely related to ANOVA (analysis of variance) and regression analysis, which also attempt to express one dependent variable as a linear combination of other features or measurements.
  - LDA has continuous independent variables and a categorical dependent variable (i.e. the class label).
  - Logistic and probit regression are similar to LDA, as they also explain a categorical variable by the values of continuous independent variables.
  - A fundamental assumption of the LDA method is that the independent variables are normally distributed. (Logistic and probit regression don't have this assumption)
- **CCR** - Continuous Complete Remission – no relapse of the cancer
- **Cox Proportional Hazard Model** - examines the relationship of the survival distribution to covariates. Most commonly, this examination entails the specification of a linear-like model for the log hazard. A parametric model based on the exponential distribution may be written as

$$\log h_i(t) = \alpha + \beta_1 x_{i1} + \beta_2 x_{ik} + \cdots + \beta_k x_{ik}$$

- **Log-rank test on Kaplan-Meier Data** - calculates the χ^2 for each event time for each group and sums the results. The summed results for each group are added to derive the ultimate χ^2 to compare the full curves of each group.