1 Microarray analysis and simulation of data (Projekt 1)

1.1 Distribution and simulation of gene expression data

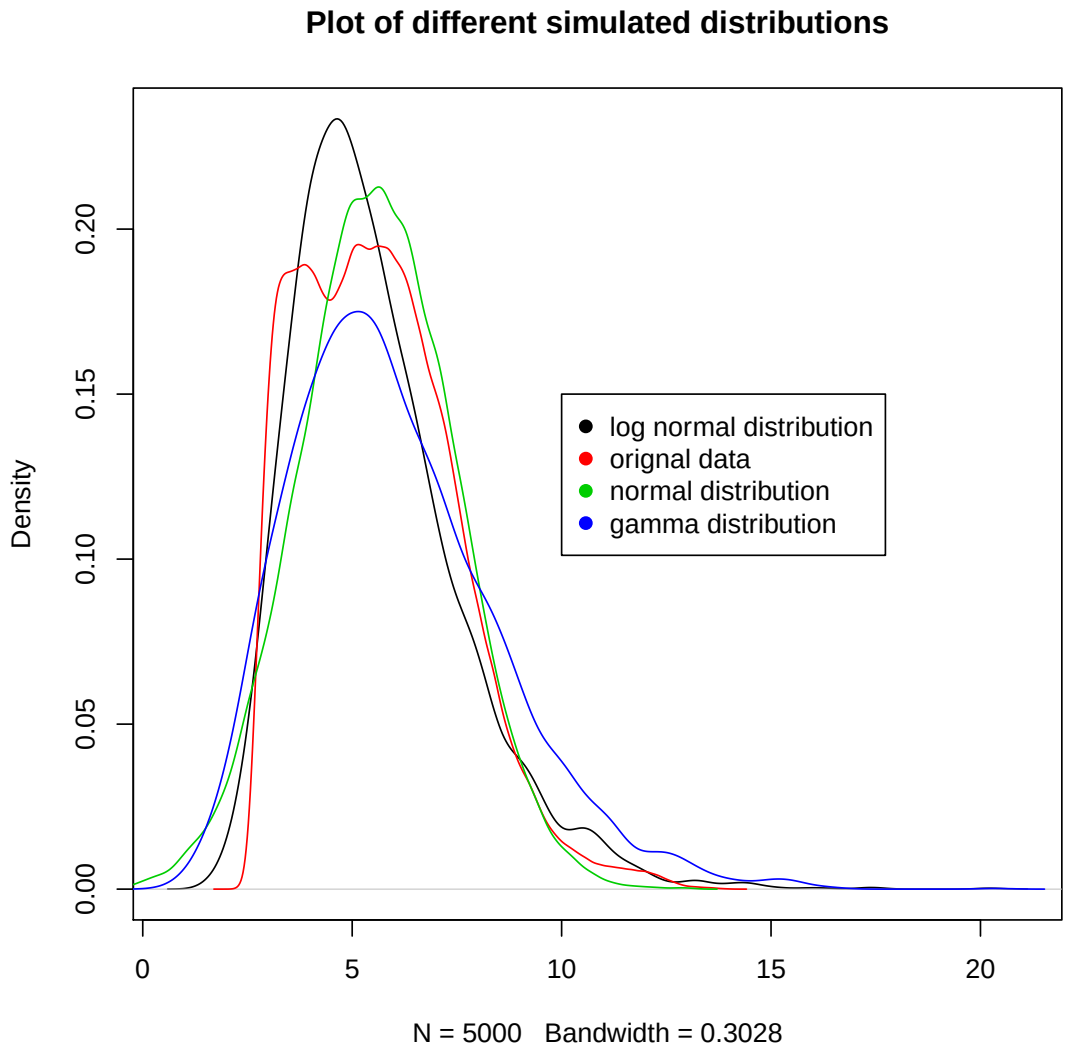

Abbildung 1: Plots der verschiedenen Verteilungen

- *Normalverteilung* ist symmetrisch und passt deswegen nicht optimal
- *Gammaverteilung* ist zu flach und breit
- *Logarithmische Normalverteilung* ist unserer Meinung nach am besten geeignet, da sie vor allem die Ränder des Graphen gut annähert.

Die Verteilung der Mittelwerte und der Varianzen der einzelnen Gene sind im Gegensatz zur Simulation in den Messwerten nicht normalverteilt. Das lässt auf eine bestimmte Ordnung schließen und zeigt, dass die Werte keine Zufallsdaten sind.

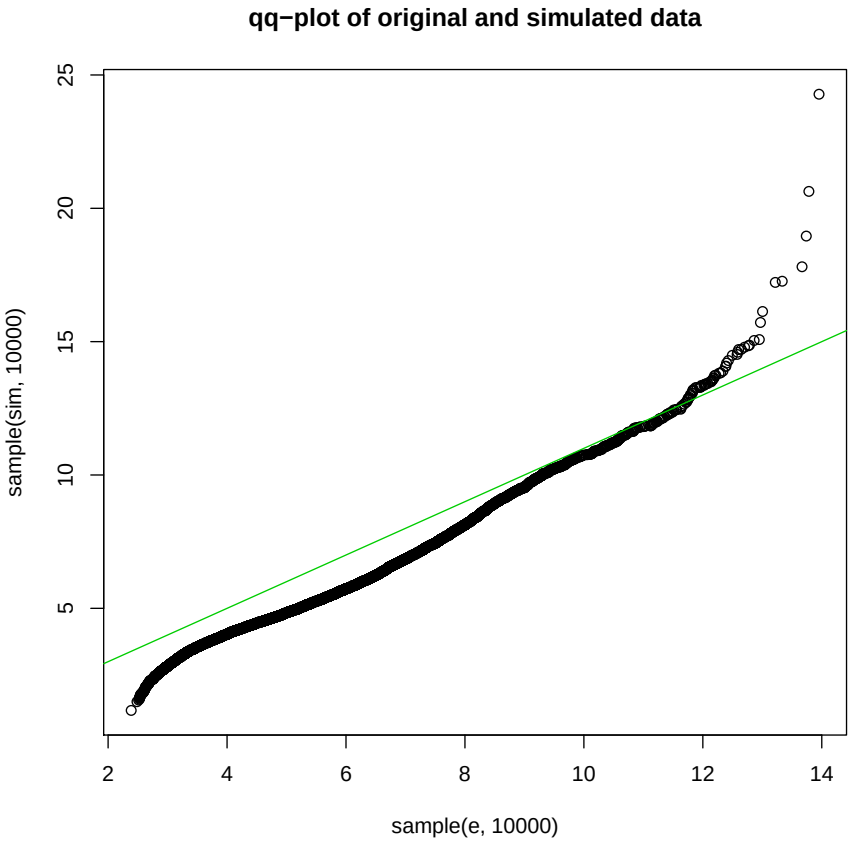

Abbildung 2: qq-Plot von den Orignialdaten gegen die simulierten Daten der logarithmischen Normalverteilung

Frage: Welcher Zusammenhang besteht zwischen den folgenden Gleichungen?

$$\cos(\alpha) = \frac{\langle x, y \rangle}{||x||||y||}$$

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 \sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

Antwort: Für den Fall, dass der Mittelwert der Daten bei 0 liegt gilt:

$$r = \frac{\sum_{i=1}^{n} x_i \; y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \; y_i^2}} = \frac{\langle x, y \rangle}{||x||||y||} = \cos(\alpha)$$

Das heißt, wenn der Mittelwert bei 0 liegt sind die beiden Gleichungen identisch.
Allgemein gehen die Einträge eines Vektor immer von Null aus. Berechnet man nun also den

Winkel zweier Vektoren, so berechnet man immer den Winkel, den die beiden Vektoren mit dem Nullpunkt bilden.
Der empirische Korrelationskoeffizient berechnet nun im Prinzip den Winkel, den die "Vektoren" (in diesem Fall die x, bzw. y- Werte) mit dem Mittelwert einschließen.
Dabei gilt:
Ein 90° Winkel ($r = 0$) zwischen den Vektoren, bedeutet, dass diese nicht korrelieren. Liegen die Vektoren dagegen übereinander (also $0° - 180°$) sind sie sehr stark korreliert. ($r = -1$ oder $r = 1$)

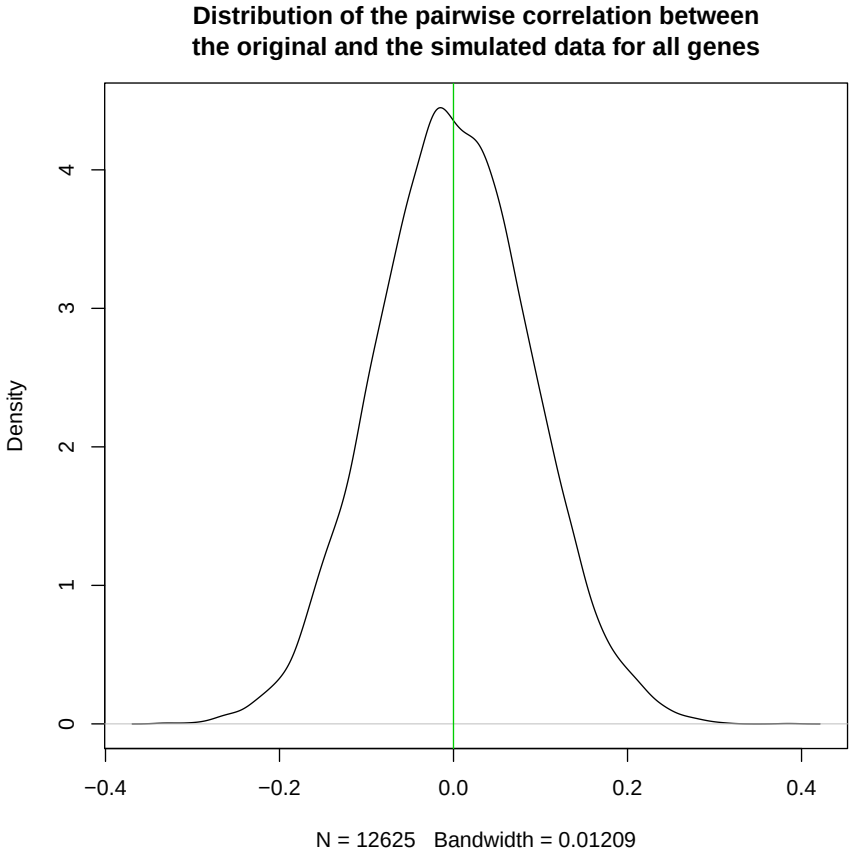

Abbildung 3: Verteilung der paarweisen Korrelation der Gene von Originaldaten und Simulation

Die paarweise Korrelation der Gene von Originaldaten und Simulation (Abbildung 3) ist normalverteilt um den Mittelwert 0; das zeigt, dass die Simulation die Originaldaten nicht nachbildet sondern nur aus Zufallswerten besteht, welche nicht mit den (nicht zufälligen) Originaldaten korrelieren.

1.2 Interpretation of gene expression values and identification of subgroups

38319_at (CD3D-Gen) ist für den Rezeptor für T-Zellen verantwortlich. Wie im Plot (Abbildung 4) ersichtlich ist bei Patienten mit T-Zellen-ALL das CD3D-Gen deutlich höher exprimiert.

38355_at (DDX3Y-Gen) ist eine RNA-Helikase; die Verteilung deutet nicht auf eine Krankheit hin, das Gen liegt auf dem Y-Chromosom und ist deswegen bei Männern deutlich höher als bei Frauen exprimiert (siehe Abbildung 5)

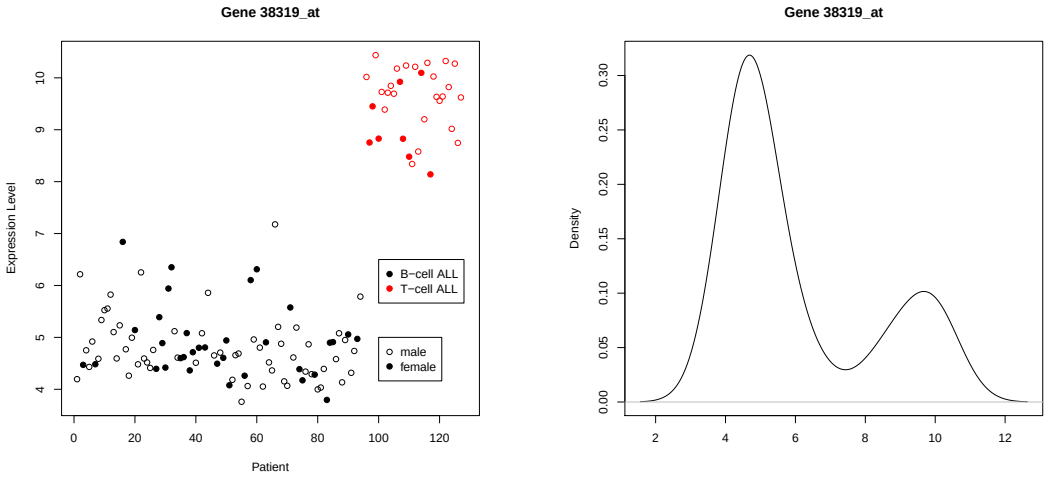

Abbildung 4: 38319_at: CD3D-Gen

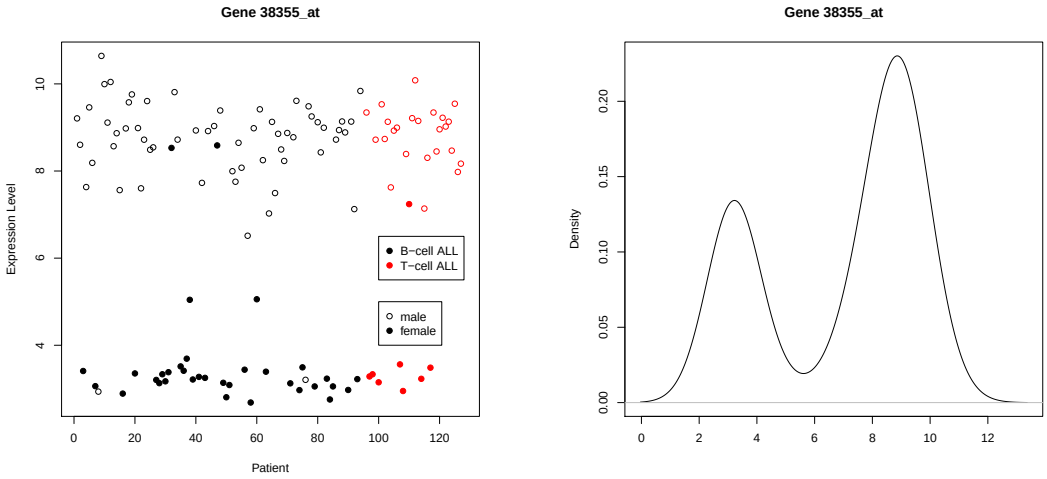

Abbildung 5: 38319_at: DDX3Y-Gen

Mit Hilfe des Korrelationskoeffizienten lassen sich ähnliche Gene finden: Man vergleicht den Korrelationskoeffizienten des Gens mit allen anderen Genen im ExpressionSet und wählt denjenigen dessen Betrag möglichst hoch ist.

Hat man ein Kriterium des Gens ausgemacht kann man auch speziell nach diesem Kriterium filtern, z.B. nach Geschlecht (wie bei 38355_at, Abbildung 6) oder Typ der Erkrankung (wie bei 38319_at, Abbildung 7). Dazu kann man die **pData()**-Funktion von Bioconductor nutzen um die Zusatzinformationen der ALL-Datensätze auszulesen und dann die ExpressionSets dementsprechend filtern.

Anhand der Varianz der Gene lassen sich auch ohne ein vorher bestimmtes Gen zum Vergleich auszuwählen “interessante” Gene finden: Gene, die eine Veränderung in den Datensätzen unterschiedlicher Patienten aufweisen haben eine höhere Varianz und werden so wenn man die Varianzen aller Gene (aufsteigend) sortiert am Ende der Liste erscheinen; dort finden sich also Gene die “interessant” für genauere Untersuchung sein können (z.B. siehe Abbildung 8)

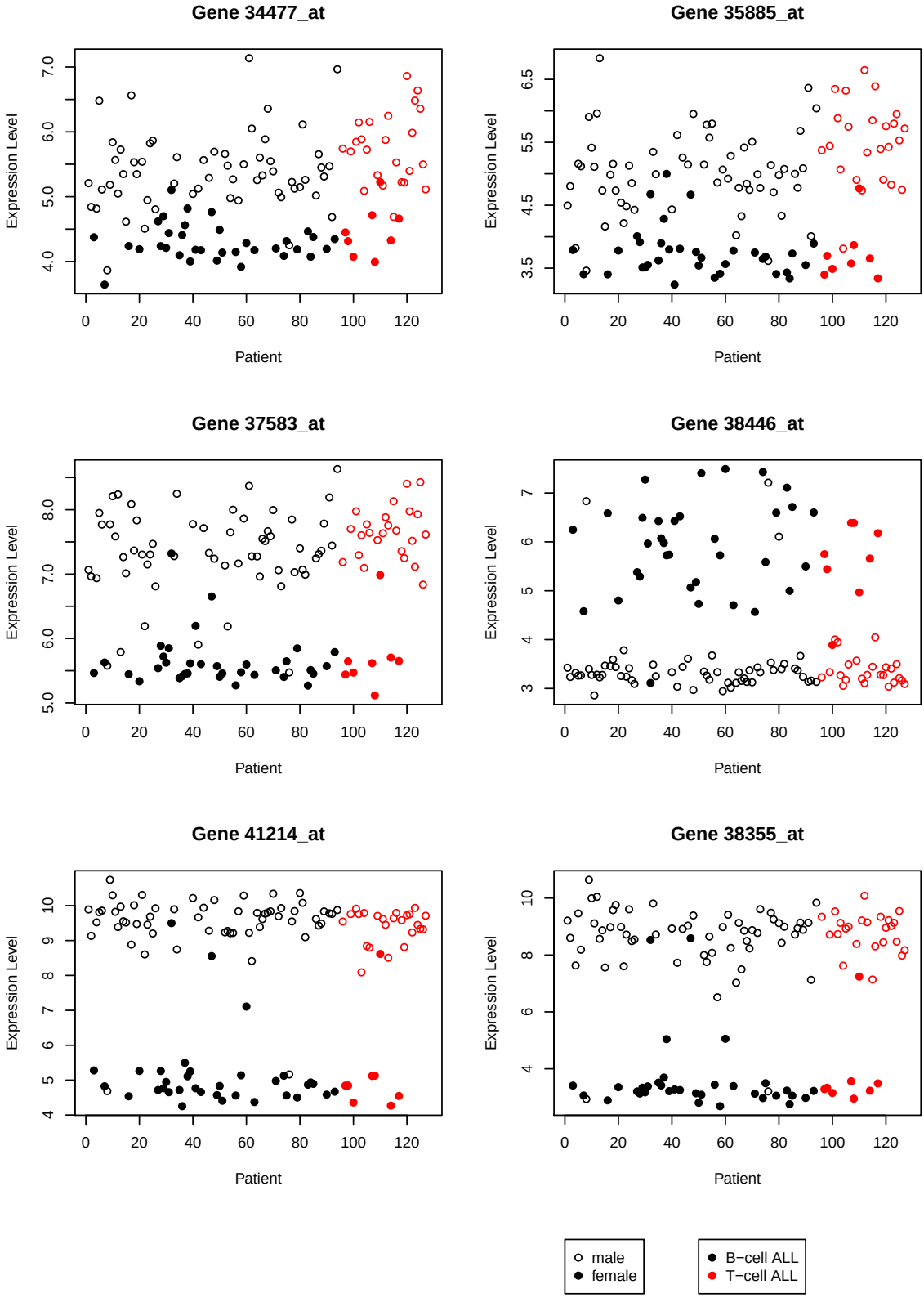

Abbildung 6: Gene, deren Expressionen sich ähnlich zum 38355_at-Gen verhalten (abhängig vom Geschlecht des Patienten)

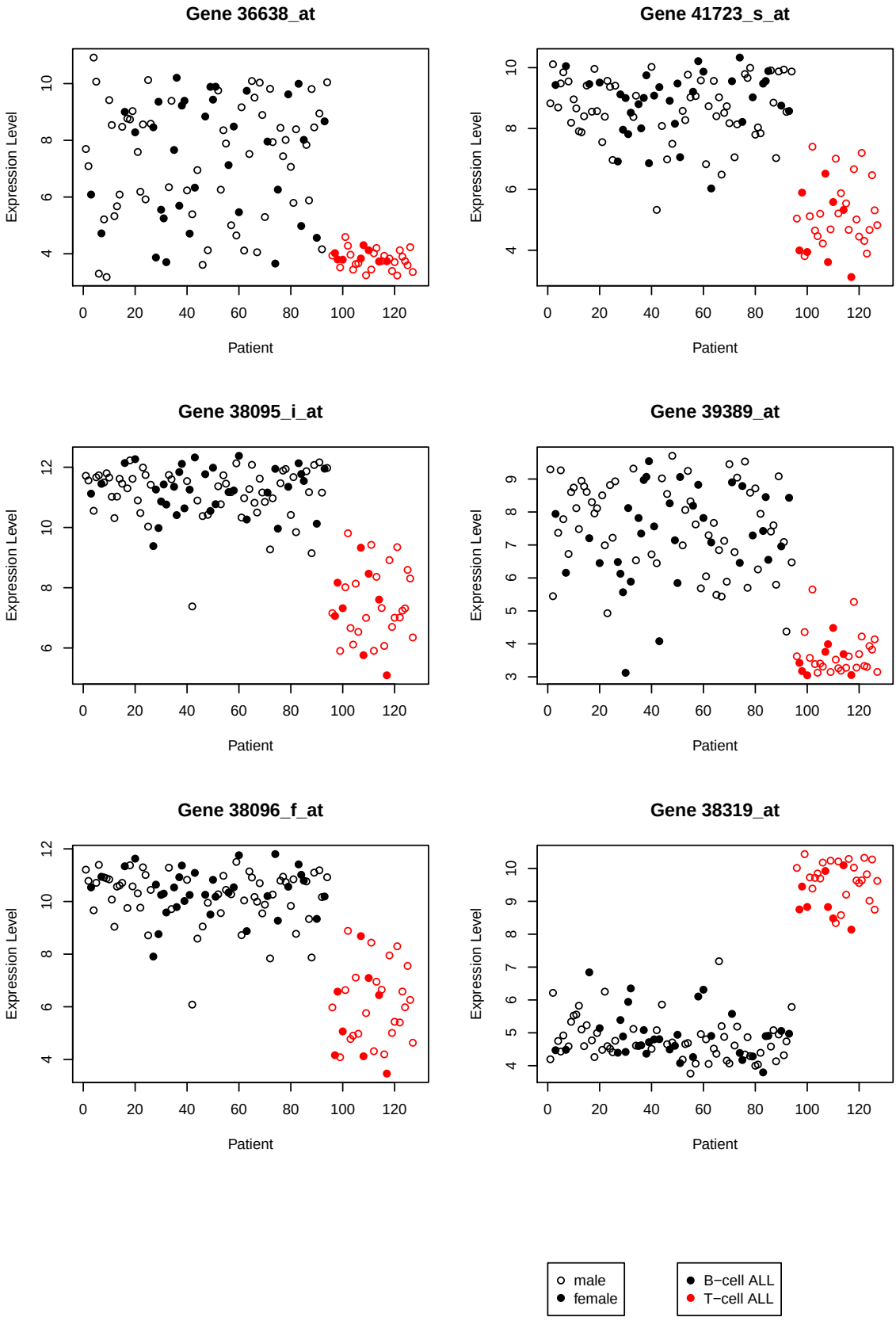

Abbildung 7: Gene, deren Expressionen sich ähnlich zum 38319_at-Gen verhalten (abhängig vom Typ der Erkrankung)

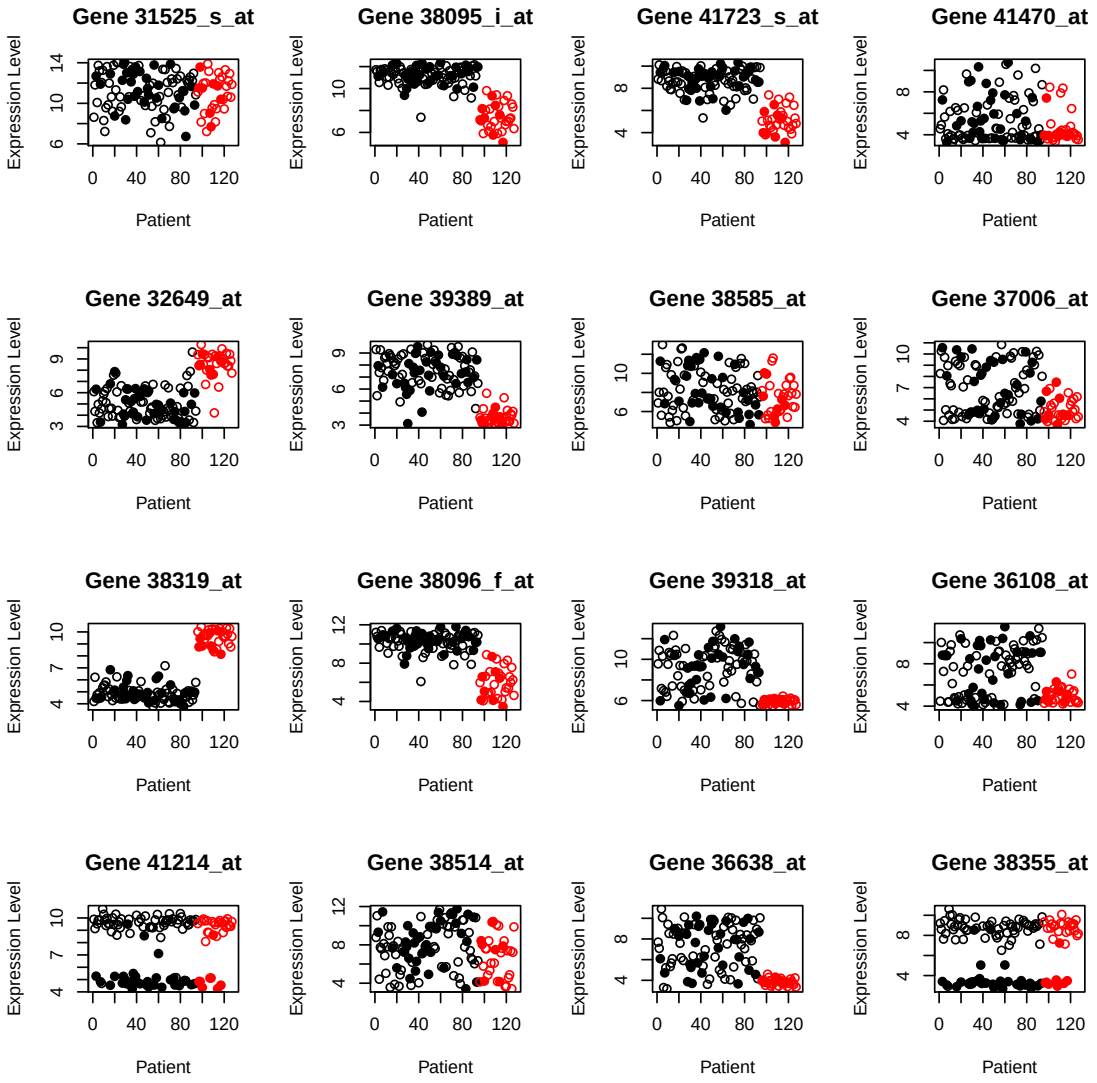

Abbildung 8: Plots der 16 Gene mit der höchsten Varianz aus dem ALL-ExpressionSet