Genomics VL. 7 bei Dr. Jung Differenzierte Expressionsanalyse

Wir haben die Hälfte fast hinter uns. In der ersten Hälfte des Moduls haben wir uns mit Genomanalyse befasst. Das heißt wir haben uns mit Analysen auf DNA-Ebene befasst. Wir schwenken jetzt um auf das Transkriptom. Das spielt auch in der Infektionsforschung eine große Rolle. Das Genom ist ja etwas Statisches, weil es sich in der Regel nicht groß verändert. Das Transkriptom hingegen kann auf äußere Einflüsse reagieren. Das bedeutet, dass in jedem Organ und in jeder Zelle etwas anderes exprimiert/ abgelesen wird. Das Transkriptom ist in der Infektionsforschung wichtig, weil eine Infektion oder eine Krankheit als externe Einflüsse auch die Expression ändern.

S.3-5. **Biologische Grundlagen**

Wie ich bereits gesagt habe, ist DNA statisch. Die Entscheidung, welches Gen zu welchem Zeitpunkt abgelesen wird und woraus ein Protein gebaut wird, ist abhängig vom jeweiligen Zustand des Gewebes. Typischerweise ist es so, dass bei der Analyse des Transkriptoms nicht nur die Genexpressionsdaten erhoben werden, sondern es werden auch viele Zusatzdaten aus dem Internet gezogen. Das kennen wir aber bereits aus der DNA-Analyse. Auch dort haben wir zum Beispiel viele Referenzsequenzen benutzt. Im Bereich der Genexpressionsanalyse brauchen wir sehr viele Annotationen für Gene. Wir brauchen also Genname, aber auch Funktionen von bestimmten Genen und zusätzliche Informationen wie zum Beispiel die Pathways, wo die Gene eine Rolle spielen. Das heißt, hier haben wir eine Situation, wo wir nicht nur mit einer Datei arbeiten können. Wir müssen viel Zusatzinformation zur Datenanalyse hinzuziehen. Transkriptomanalyse kann man als kontrolliertes Experiment im Labor machen. Wir können zum Beispiel Zelllinien analysieren und dann vergleichen wir zum Beispiel in einem Zweigruppen-Experiment Kontrollzelllinien und Zelllinien, die wir mit etwas behandelt haben (mit etwas toxischem oder einem Virus zum Beispiel). Wir vergleichen Kontrollgruppe und behandelte Gruppe und schauen uns an, wie sich die Expression von bestimmten Genen zwischen beiden Gruppen verändert hat. Dann haben wir ein kontrolliertes Experiment. Das Transkriptom wird aber auch oft am lebenden Organismus analysiert. Dazu nimmt man zum Beispiel Blutproben oder Gewebeproben. Im Bereich der Infektionsforschung würde man eher eine Blutprobe nehmen. Im Bereich der Krebsforschung nimmt man dann eine Tumorbiopsie und schaut sich an, wie sich die Genexpression der Tumorzellen im Rahmen der Krebserkrankung verändert hat im Vergleich zu normalen Körperzellen. Wenn man in der Krebsforschung die Lebenszeitdaten hat (wie lange nach der Diagnose hat ein Patient überlebt oder wie lange nach der Diagnose trat ein bestimmtes Event ein wie zum Beispiel Metastasierungen) oder Information über die Medikamente, die vom Patienten genutzt wurden, dann kann man all diese Informationen über den Patienten mit seinem Transkriptom in Verbindung setzen. Wir haben also eine Situation, die wir als Datenwissenschaftler als eine “BID DATA Situation” bezeichnen. Auf der DNA-Ebene hatten wir Referenzgenome und Annotationen, die wir benutzt haben. Hier haben wir viel mehr Annotationen, die wir aus dem Internet beziehen und wir müssen Patientendaten heranziehen. Es kann aber sein, dass wir es mit kontrollierten Experimenten zu tun haben. Dann haben wir nicht so viele Annotationen.

s.6. Transkriptom.

Beim Transkriptom müssen wir nochmal aufpassen. Wir wissen ja bereits, dass es nicht so ist, dass ein bestimmter Abschnitt auf dem Genom für ein bestimmtes Gen codiert, sondern **ein Gen setzt sich in der Regel aus mehreren Abschnitten (Exons) zusammen**. Diese Abschnitte können

unterschiedlich gespliced werden. Hier hängt es also davon ab, mit welcher Technik ich die Genexpression erhebe. Wenn ich sogenannte Microarrays verwende, dann habe ich nur eine fixe Anzahl an Splicevarianten für die ich meine Genexpression messe. Der Vorteil von RNA-Sequenzierung mit NGS (next generation sequencing) ist, dass ich auch Splicevarianten entdecken kann, die noch nicht bekannt sind. Ich kann also noch unbekannte Zusammensetzungen von Exonen, die zu einem Gen führen, entdecken.

s.7. **RNA-Seq: Messen von Genexpression mit NGS.**

Schauen wir uns nochmal an, wie wir die Genexpression mit NGS messen können. Wir brauchen auch hier ein **Referenzgenom**, des Organismus an dem die Genexpression erhoben wird. Wir mappen die Sequenzier-Reads gegen das Referenzgenom mit den bereits bekannten Methoden. Jetzt geht es aber nicht darum bei den angemappten Reads nach Mutationen zu suchen –also Unterschiede zwischen zum Referenzgenom, sondern wir wollen auszählen, wie viele Reads an ein bestimmtes Gen mappen. Nehmen wir zum Beispiel an, dass wir ein Referenzgenom mit zwei Genen draufhaben. Auf dem ersten Gen mappen 7 Reads an. An das zweite Gen mappen 9 Reads. Man würde jetzt schlussfolgern, dass vom zweiten Gen mehr abgelesen wird. Das zweite Gen war also höher exprimiert. Aus RNA-Seq erhalten wir so genannte "**Count-Daten**" - also Zähldaten. Count-Daten, da wir darüber zählen, wie viele Reads an ein bestimmtes Gen mappen. Diese Count-Daten sind dann **unser Expressionswert**. Das spielt bei der statistischen Analyse dieser Daten eine Rolle. Die Statistik befasst sich ja mit bestimmten Wahrscheinlichkeitsverteilungen und wenn ich Zählwerte habe wie diese, dann habe ich Ganzzahlige Informationen. Ich kann ja keine Zahl mit Kommas bekommen, da mein Reads entweder bindet oder nicht. Ich kriege also immer Count-Daten, die aus ganzen Zahlen bestehen. Wir werden später bei der Auswertung sehen, dass es man deshalb eine andere Methode verwendet als bei Werten die aus Dezimalzahlen bestehen.

s.8. **Messen von Genexpression mit DNA-Microarray.**

Die DNA-Microarrays habe ich schon mal als Begriff gebracht. Microarray wurde Mitte der 90er Jahre erfunden. Die Idee ist, dass man auf einem kleinen Chip – Microarray genannt- komplementäre DNA in Spots aufträgt und zwar kriegt jedes Gen einen Spot. Deshalb heißt es auch DNA-Microarray, weil man einzelsträngige DNA, die komplementär zu mRNA ist, nutzt. Ich nehme quasi von jedem Gen die einzelsträngige DNA und trage Sie auf einem Array auf. So ein Array ist vielleicht 2*2 cm groß und hat viele Punkte, wo die DNAs aufgetragen sind, und man kann es dann unterm Mikroskop analysieren. Wie messe ich aber jetzt die Genexpression? Ich nehme aus meiner Gewebeprobe. Aus der isoliere ich mRNA. Ich habe damit exprimierten Gene zu dem Zeitpunkt isoliert. Diese mRNAs markiere ich jetzt mit einem fluoreszierenden Farbstoff. Dann wasche ich meine markierte mRNA auf den Chip drauf. Sie können sich schon denken, was als Nächstes passiert: die mRNAs binden an ihre komplementäre DNA in einem der Spots auf dem Array-Chip. Das heißt, die mRNA bleibt an einer bestimmten Position an der passenden DNA hängen. Wenn das Gen hoch exprimiert war, dann leuchtet der entsprechende Spot. Gene, die keine mRNA exprimiert haben, leuchten nicht und der Spot bleibt schwarz. Diese Fluoreszenz kann ich (in Form von Dezimalzahlen und nicht ganzen Zahlen) messen. Ich kriege einen Fluoreszenzwert. Hier sollte es eindeutig werden, dass ich unterschiedliche statistische Methoden brauche, wenn ich ganzzahlige Count-Daten in RNA-Seq mit NGS auswerte im Vergleich zu Microarrays, wo ich Fluoreszenzwerte, die nicht Ganzzahlig sind, auswerte. Und ich betone nochmal: Der Vorteil von RNA-Seq ist, dass man neue Spilcevarianten entdecken kann. Der Microarray ist vorpräpariert mit einer fixen Anzahl an Genen in Form einzelsträngiger DNA. Microarrays werden aber noch heute benutzt, weil Sie noch günstiger sind als RNA-Seq.

s.9. Messen von Proteinexpression.

Analyse von Genexpressionsdaten lässt sich fast 1:1 auf Proteinexpressionsdaten übertragen. Wir haben ja neben dem Transkriptom noch das Proteom als weitere Stufe. Die Daten sehr relativ ähnlich aus. Man hat auch eine ähnliche Fragestellung: man möchte nach Proteinen schauen, die unter unterschiedlichen Zuständen und in unterschiedlichen Geweben in unterschiedlichen Mengen gebaut werden im Gewebe. Vor ein paar Jahrzehnten nahm man noch an, dass jedes Gen, welches abgelesen wird, für den Aufbau von einem Protein verantwortlich ist. Das ist das Dogma " 1 Gen = 1 Protein". Inzwischen wir, dass es nicht der Fall ist, weil sich ein Protein nach Aufbau auch unterschiedlich falten kann und durch Splicevarianten kann man aus mehreren Exon-Zusammensetzungen mehrere Proteine generieren und nicht nur 1 Protein. Insofern ist die Menge an Proteinen auch deutlich größer als die Menge an Genen. Trotzdem ist die Fragestellung ähnlich und die Methodik der Datenauswertung, die wir heute kennenlernen werden, lässt sich auf Transkriptome und Proteine anwenden. Man hat jetzt aber etwas andere Messmethoden bei den Proteinen. Zum einem nutzt man bei Proteinen die Massenspektrometrie. Dabei wird eine Proteinprobe in Peptide zerteilt. Es gibt verschiedene Varianten der Massenspektometrie, aber normalerweise nutzt man Time of Flight Massenspektometer. Da werden die Peptide durch eine Röhre geschossen und man beobachtet, wie lange ein Molekül braucht, um durch die Röhre zu fliegen. Die Peptide werden am Ende der Röhre durch einen Detektor detektiert und schwere Moleküle brauchen entsprechend länger. Über das Masse/Ladungs-Verhältnis kann bestimmen, um welches Protein es sich gehandelt hat. So kann man später prüfen, wie viel von einem bestimmten Protein in der Probe war. Eine andere Möglichkeit ist die 2D-Gelelektrophorese. Hierbei werden die Proteine in einem zwei-dimensionalen Gel nach Masse und Ladung aufgetrennt. Vorher werden Sie aber eingefärbt - entweder durch eine Silberfärbung oder durch einen anderen fluoreszierenden Farbstoff. Dann kann ich auch Proteinexpressionen messen. Im Gel ist es so, dass je größer ein bestimmter Spot /Fleck ist, desto mehr wurde von dem Protein in der Zelle gebildet. Die Größe eines Spots spiegelt also nicht das Gewicht oder Größe eines Proteins wider, sondern es zeigt wie viel vom Protein gebildet wurde. Bei Proteinen spricht man nicht von "Expression" wie bei Genen, sondern man sagt, wie viel vom Protein gebildet wurden ist.

s.10. Zwei-Gruppen Versuchsdesign.

Dann können wir uns an der Stelle schon mal anschauen, wie die Daten in so einem typischen Experiment ausschauen. Es ist etwas bequemer als bei DNA-Sequenzierungen, wo wir sehr lange Text-Files haben. Die Text-Files erhalten wir hier zwar auch als Zwischenstufe und wir kriegen hier auch .sam-files, wenn wir RNA-seq verwenden, worüber wir dann sehen können, welcher Read an welches Gen gemappt hat. Nach weiter Datenprozessierung erhalten wir dann aber eine Tabelle wie man es in Abbildung 1 sehen kann. Die meisten Experimente oder Studien haben ein so genanntes Zwei-Gruppen Design. Man hat zum Beispiel eine Kontrollgruppe und man vergleicht diese Gruppe mit einer Gruppe erkrankter Individuen. Alternativ hat man beim kontrollierten Experiment behandelte Proben in der einen Gruppe und unbehandelte Proben in der zweiten Gruppe. "n_1" und "n_2" stehen hier für die Stichprobenumfänge. Das ist natürlich auch ein Grundprinzip in der Statistik: in einem Experiment braucht man immer unabhängige biologische Wiederholungen. Das brauch ich, denn es gibt natürlich Unterschiede zwischen den Individuen. Jeder Mensch reagiert zum Beispiel unterschiedlich auf eine Infektion oder Erkrankung. Deshalb kann ich nicht nur einen Gesunden und einen Erkrankten miteinander vergleichen und daraus eine wissenschaftliche Schlussfolgerung ziehen, sondern ich brauche unabhängige Wiederholungen. Nehmen wir jetzt an, wie haben "n1"-Personen in der Kontrollgruppe und "n2"-Personen unter den Erkrankten. Normalerweise hat man

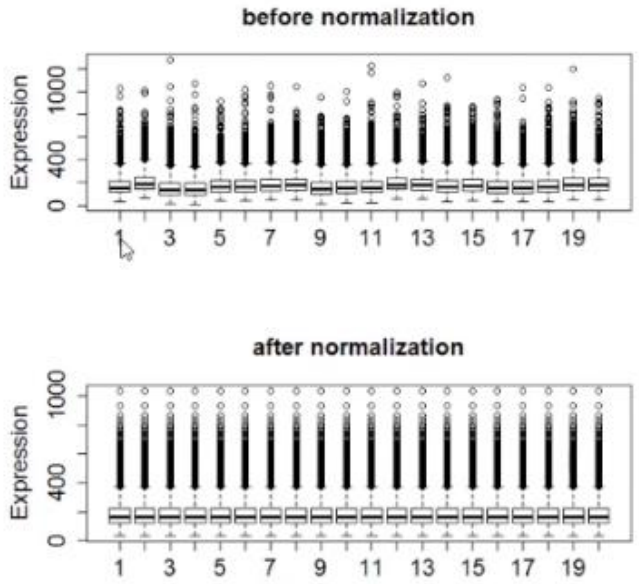
eine überschaubare Anzahl an Spalten im Datensatz. Es gibt zwar Studien, die Hunderte Patienten betrachten, aber bei einem kontrollierten Experiment mit Zelllinien kann es auch sein, dass man nur 3 Kontrollproben hat und 3 behandelte Proben. Es gibt also auch sehr kleine Experimente. Die Anzahl an Zeilen beträgt jedoch mehrere Tausend. Beim Menschen haben wir um die 20.000 Gene. Bei anderen Tieren mag die Zahl nicht so hoch sein, aber trotzdem wird man Tausende Zeilen haben. Das heißt, jede Zeile steht für ein Gen. Ich notiere für die Anzahl an Genen ein kleines “d” für “dimension”. Das kommt aus der statistischen Notation. Dort notiert man die Anzahl an Merkmalen mit einem kleinen d. Vielleicht kann ich an der Stelle schon erwähnen, dass die klassische Statistik, die in der ersten Hälfte des 20. Jahrhunderts entwickelt wurden ist, Niedrigdimensionale Daten betrachtet. Das heißt, man erwartet hier, dass man sehr viele Samples hat – also hätte man mehr Spalten- und wenige Merkmale – also wenige Zeilen. So war es auch früher: damals haben sich Arbeitsgruppen mit wenigen Genen befasst. Durch die Hochdurchsatzverfahren (microarrays, NGS) kann man jetzt Tausende Gene gleichzeitig betrachten. Dann erhält man einen Datensatz, den man in der Statistik als Hochdimensionale Daten bezeichnet. Also jedes gen betrachten wir hier quasi als Merkmal und wir haben ja mehrere Tausende Merkmale/Gene. Die Anzahl an Merkmalen ist um ein Vielfaches größer als der Stichprobenumfang (die Anzahl an Samples). Das versteht man unter Hochdimensionalen Daten. Das ist die Maßgebende Struktur, die wir in der Bioinformatik betrachten. Das ist auch der Unterschied zwischen Bioinformatik und Biostatistik, wo man Niedrigdimensionale Datensätze hat. Was habe ich dann in den Tabellen drinnen stehen? Das sind dann die Expressionswerte. Also wie viel wurde von Gen 1 in Sample 1 exprimiert in Abbildung 1? Da steht ein Wert von 7.6. Das könnte zum Beispiel der Fluoreszenzwert von einem Microarray sein. Bei RNA-Seq, wo wir Count-Daten haben, haben wir keine Dezimalzahlen, sondern ganze Zahlenwerte in der Tabelle stehen. Ziel der Analyse ist es Gene zu finden, die unterschiedliche Expressionen zwischen den zwei Gruppen aufweisen.

Abbildung 1: Zwei Gruppen Versuchsdesign- Tabelle.

Gen	Kontrollgruppe			Erkrankte		
	1	…	n_1	1	…	n_2
1	7.6	…	7.5	4.2	…	5.1
2	5.0	…	4.9	8.2	…	7.3
3	3.9	…	4.2	8.2	…	7.5
4	5.9	…	6.2	1.8	…	1.0
5	5.9	…	5.9	1.3	…	2.3
…	…	…	…	…	…	…
d	9.4	…	9.3	8.7	…	9.2

s.11. Datenvorverarbeitung

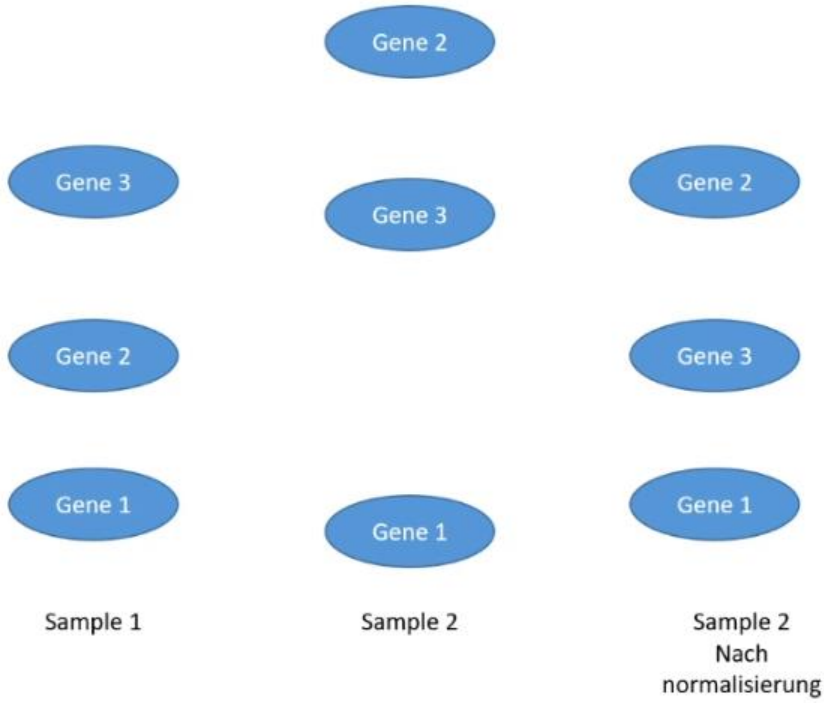
Abbildung 2: Microrray Datenvorverarbeitung



Wir wollen uns jetzt erstmal anschauen, wie man die Daten weiter aufarbeiten muss, bevor wir mit der eigentlichen Analyse starten können. Die Analyse würde darin bestehen, dass man schaut, ob ein Gen bei den Erkrankten deutlich mehr oder weniger exprimiert wird als bei den Gesunden Samples. Es sind aber eine Reihe an Vorverarbeitungsschritten notwendig, bevor wir mit der differenziellen Expressionsanalyse starten können. Insbesondere bei den **Microarrays** ist es so, dass ich für jedes Sample ein eigenes Microarray nutze und dabei kann sich immer ein kleiner technischer Fehler einschleichen. Vielleicht habe ich 2 Proben mit je einem Microarray und in der einen Probe habe ich dann systematisch dann immer höhere Werte gemessen als bei dem Anderen. Das ist ein Laboreffekt, den ich hinterher rausrechnen muss. Schauen wir und die obere Skizze auf Abbildung 2 an. Wir gehen davon aus, dass wir 20 Samples haben (X-Achse). Pro Sample haben wir dann ein Microarray oder einen RNA-Seq-Lauf, denn bei RNA-Seq haben wir auch technische Fehler. Auf der Y-Achse haben wir dann die Genexpression und pro Sample haben wir da ein Boxplot. Boxplots gehen immer vom Minimum bis zum Maximum. Darüber werden noch Extremwerte eingezeichnet. Der Dicke Balken innerhalb der Box notiert den Median. Nehmen wir zum Beispiel den Menschen, dann steckt hinter jedem Boxplot 20.000 Gene und ihre Expression. Der Boxplot zeigt Gene an, die hoch exprimiert werden und niedrig exprimiert werden. Wenn wir jetzt in der oberen Skizze Samples 1 bis 20 vergleichen, dann merken wir, dass Sie nicht alle gleich liegen. Die Verteilungen wackeln ein bisschen und das ist eben dieser technische Fehler. Wir gehen natürlich auch davon aus, dass wir einen biologischen Unterschied haben. Das bedeutet, wir erwarten einen biologischen – und einen technischen Effekt. Beim biologischen Effekt gehen wir in der Regel davon aus, dass er eher klein ist. Wir gehen also davon aus, dass eine überschaubare Anzahl an Genen durch eine Erkrankung zum Beispiel unterschiedliche Expressionsprofile haben. Wenn wir davon ausgehen, dass der biologische Effekt nicht so groß ist, dann müssen wir davon ausgehen, dass die beobachteten Schwankungen in der oberen Skizze von Abbildung 2 auf technische Fehler zurückzuführen sind. Wir versuchen diesen Fehler zu korrigieren. Das macht man zum Beispiel mit der **Quantil-Normalisierung**. Bei der Quantil-

Normalisierung werden alle Daten der Chips auf das gleiche Level geschoben. Man sieht auf der unteren Skizze der Abbildung 2 wie die Boxplots nach der Normalisierung aussehen. Nachdem die Schwankungen behoben wurden, kann ich davon ausgehen, dass kein technischer Fehler vorliegt. Die Unterschiede, die übrig bleiben, sind jetzt nur biologischer Natur. Ich will das nochmal für Sie skizieren (Abbildung 3).

Abbildung 3: Skizze.

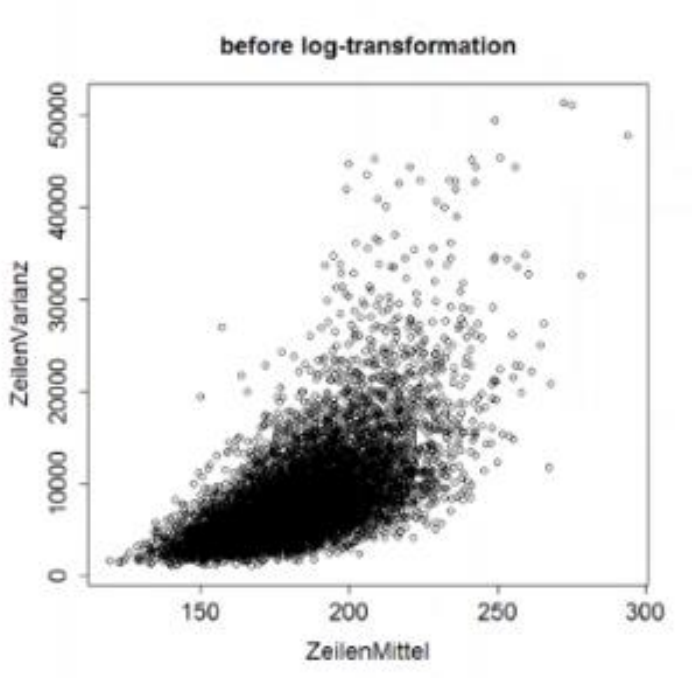


Nehmen wir mal an, wir haben mit einem Microarray oder RNA-Seq die Expression von 3 Genen gelesen (Abbildung 3). Gen 1 hat die niedrigste Expression und liegt deshalb ganz unten. Gen 2 hat die zweithöchste Expression und Gen 3 hat die höchste Expression. Von Gen 3 wurde also am Meisten von der DNA abgelesen. Nehmen wir mal an, dass ist die Expression dieser drei Gene in Sample 1. Jetzt haben wir aber mehr als ein Individuum, aus dem wir Samples beziehen und analysieren. Sample 1 kommt von einem gesunden Individuum und Sample 2 kommt von einem Erkranktem. Sagen wir mal in Sample 2 ist Gen 1 immer noch am Niedrigsten Exprimiert, aber Gen 2 ist durch die Erkrankung deutlich höher exprimiert als Gen 1 oder 3. gen 3 ist immer noch hoch exprimiert, aber nicht so hoch wie im gesunden Individuum. Jetzt kann es vor der Normalisierung sein, dass ich nicht nur einen biologischen Effekt vorliegen habe, sondern auch einen technischen Effekt. Das behebe ich durch die Quantil-Normalisierung. Die Normalisierung soll alle so verschieben, dass sie gleich sind. Das heißt Gen 1 von Sample 2 wird auf die Höhe von Sample 1 gehoben. Das zweithöchste exprimierte Gen in Sample 2 – Gen 3- schiebe ich auf die Höhe des zweithöchst exprimierten Gens von Sample 1 – Gen 2. Gen 2 von Sample 2 schiebe ich auch runter auf die Höhe von Gen 3 von Sample 1. Wir sehen, dass, nachdem Sample 2 normalisiert wurden ist, wir dieselben Messwerte in Sample 2 und 1 haben, aber wir haben eine andere Reihenfolge der Gene. Sample 1 hat die Reihenfolge 1,2,3 und Sample 2 hat die Reigenfolge 1,3,2. Wir haben also dieselben Messwerte aber eine andere Reihenfolge. Das wiederum heißt, dass der biologische Effekt erhalten bleibt. Den technischen Fehler, der dadurch entsteht, dass jede Probe auf einem eigenen Chip läuft

oder einen eigenen Lauf bekommt, haben wir damit aber behoben. Das ist das Prinzip der Quantil-Normalisierung. So macht man die einzelnen Proben vergleichbar.

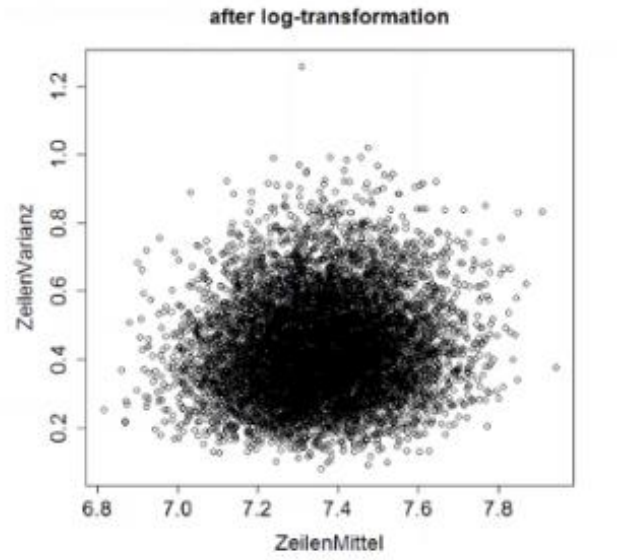
S.12-13. Datenvorverarbeitung

Abbildung 4: Varianz eines hoch exprimierten Gens.



Das ist der erste Vorverarbeitungsschritt. Ich muss noch einen zweiten Schritt machen. Der beruht darauf, dass Gene, die hoch exprimiert sind, in der Regel auch eine höhere Varianz haben. Die rohen Expressionswerte weißen in der Regel auch eine Beziehung zwischen mittlerer Expression und Varianz eines Gens auf. Wir sehen in Abbildung 4 40.000 Gene (Punkte) dargestellt. Auf der X-Achse haben wir die mittlere Genexpression und auf der Y-Achse haben wir die Varianz der einzelnen Gene. Wir sehen in Abbildung 4, dass Gene die einen höheren mittleren Expressionswert aufweisen auch eine höhere Varianz besitzen. Das ist statistisch schlecht, denn eine höhere Varianz bedeutet auch eine höhere Schwankung der Messwerte und damit auch eine höhere Unsicherheit. Ich kann dadurch nicht so gut vergleichen, ob Gene zwischen Kontrollgruppe und erkrankter Gruppe unterschiedlich stark exprimiert sind, weil ich dann alle Gene mit unterschiedlichen Maßstäben vergleiche. Einem Gen, welches an sich niedrig exprimiert ist, messe ich eine höhere Sicherheit bei, weil die Messwerte dann nicht so stark schwanken wie bei den hoch exprimierten Genen. Wir haben also eine Beziehung zwischen mittlerer Expression und der Varianz des jeweiligen Gens. Um hoch exprimierte Gene und niedrig exprimierte Gene vergleichbar zu machen, muss ich diese Varianz und diese Beziehung zu der mittleren Expression rausnehmen. Ich mache also eine **Varianzstabilisierung**. Das mache ich einfach, indem ich die Daten **logarithmiere**. Ich nehme also die Quantil normalisierten Daten aus Abbildung 1 und logarithmiere diese. Wir sehen dann in Abbildung 5, dass es diese lineare Beziehung zwischen Varianz und mittlerer Expression nicht mehr gibt. Es gibt zwar immer noch Gene, die eine höhere Varianz haben als andere, aber ist nicht mehr davon abhängig, ob ein Gen eine hohe oder niedrige Expression hat.

Abbildung 5: Abbildung 4 Nach Varianzstabilisierung.



Wir haben jetzt zwei entscheidende Schritte der Datenvorverarbeitung kennengelernt: **Varianzstabilisierung und Quantil-Normalisierung.**

S.14. Zwei-Gruppen Versuchsdesign.

Jetzt haben wir die Datenvorverarbeitung abgeschlossen und wir können jetzt mit der eigentlichen Analyse beginnen. Wir versuchen jetzt diejenigen Gene rauszufiltern, die zwischen der Kontrollgruppe und den Erkrankten unterschiedliche Expressionsprofile aufweisen. Dafür gibt es verschiedene Techniken. Bevor wir aber jetzt wirklich in die Datenanalyse einsteigen, möchte ich über Datenexploration sprechen.

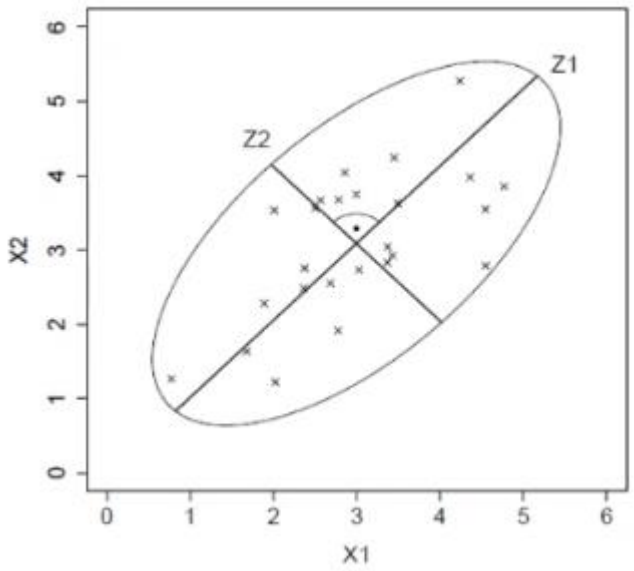
s.15. Zwei-Gruppen Versuchsdesign – Daten Exploration.

Bevor man mit der eigentlichen bioinformatischen Analyse beginnt, untersucht man die Daten daraufhin, ob es generell einen Effekt gibt zwischen Kontrollgruppe und erkrankter Gruppe zum Beispiel. Wir befinden uns also noch nicht auf Einzelgen-Ebene, sondern wir schauen uns den gesamten Datensatz an. Nun haben wir aber das Problem, dass wir es mit Hochdimensionalen Daten zu tun haben. Jetzt leben wir aber in der 3 vielleicht 4-dimensionalen Welt. Geometrisch können wir uns nur 3 Dimensionen vorstellen und wir können Daten nur in 2 oder 3-D darstellen. Wenn ich jetzt 40.000 Gene habe, dann müsste ich ein Diagramm zeichnen, das 40.000 Achsen hat. Ein Diagramm ist aber eine Darstellung auf 2 Dimensionen, nicht 40000. Es gibt jetzt aber einen sehr schönen Weg auch Hochdimensionale Daten im 2 oder 3-dimensionalen darzustellen. Das Verfahren dazu wenden die meisten von ihnen regelmäßig an, wenn Sie mit ihrem Handy ein Foto machen. Sagen wir mal, sie nehmen ihr Handy und fotografieren eine 3-dimensionale Landschaft oder einen 3-D Kopf. Dann wird doch die 3-dimensionale Information auf den 2-dimensionalen Screen ihres Handys projiziert. Das heißt auch hier machen Sie eine **Dimensionsreduktion**. Trotzdem können Sie auf dem 2-dimensionalem Bildschirm immer noch das 3-dimensionale Objekt erkennen. Das Gleiche kann man in der Mathematik mit Hochdimensionalen Räumen machen. Ich kann zum Beispiel einen 40-dimensionalen Raum auf 2 bis 3 Dimensionen runterbrechen und kann immer noch sehen, ob es

irgendwelche Effekte gibt. Ich kann also Hochdimensionale Informationen auf 2 oder 3 D runterbrechen und immer noch die Information ablesen und erkennen.

s.16. **Zwei-Gruppen Versuchsdesign – Daten Exploration.**

Abbildung 6. PCA.



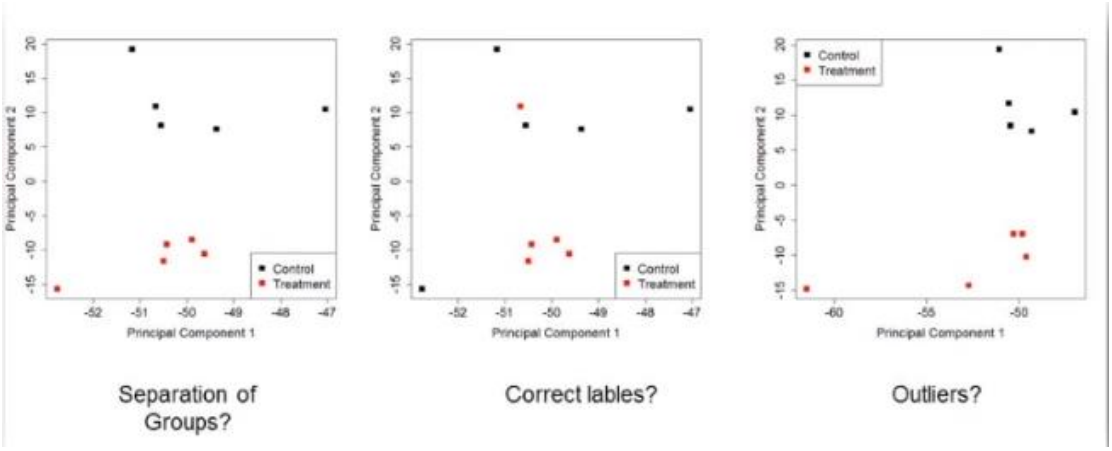
Die Methode, die man anwenden, um so einen Hochdimensionalen Datensatz auf 2 oder 3 D runter zu projizieren, nennt sich **Hauptkomponentenanalyse**. Auf Englisch nennt man es **PCA = Principal Component Analysis**. In der Literatur finden Sie sehr häufig den Begriff PCA. Ich merke in Gesprächen oft, dass Leute PCA und PCR verwechseln: Tun Sie das nicht. Ich will Ihnen ganz kurz skizieren, wie diese Hauptkomponentenanalyse funktioniert. Wir sehen auf Abbildung 6 ein Schaubild mit zwei Achsen – X1 und X2. Machen wir es uns einfach. Sagen wir mal, wir haben ein Genexpressionsexperiment, wo wir nur zwei Gene betrachten. Gen 1 liegt auf der X-Achse (X1) und Gen 2 liegt auf der Y-Achse (X2) und ich habe 20 Samples und deshalb haben wir 20 Kreuze. Wir sehen dann zum Beispiel, dass sowohl Gen 1 als auch Gen 2 in einigen Samples hoch exprimiert waren. Dann gibt es aber auch Samples, wo beide Gene niedrig exprimiert waren. Typischerweise kann ich um eine Punktwolke eine Ellipse drumlegen oder wenn beide Gene nicht miteinander korreliert sind, dann ist es eher ein Kreis. So eine Ellipse hat immer eine Achse der größten Ausdehnung. Also gibt es immer den längsten Weg von einem Rand der Ellipse zum anderen Rand der Ellipse. Das ist in Abbildung 6 die Achse Z1. Das bezeichnet man als die **Hauptachse**. Im rechten Winkel dazu liegt dann die **zweite Hauptachse**. Die Hauptkomponentenanalyse projiziert jetzt eine Punktwolke in dieses neue Koordinatensystem, welches sich über die Ellipsen-Achsen aufspannt. Die Daten werden jetzt also in den Ellipsenraum reinprojiziert und wir haben damit ein neues Koordinatensystem. Wir haben es uns jetzt einfach gemacht, weil wir mit zwei Genen nur 2 Dimensionen haben und wir projizieren es jetzt in einen anderen 2-dimensionalen Raum (Ellipsen-Achsen). Bei Genexpressionsdaten fange ich an in 40 Dimensionen, bilde dort auch ein Ellipsoid und verwende von diesem Ellipsoid am Schluss nur noch die erste und zweite Hauptachse.

s.17. **Zwei-Gruppen Versuchsdesign – Daten Exploration.**

So, wie schaut es hinterher aus? Wenn ich meine Daten in diesen neuen niedrigdimensionalen Raum projiziert habe, kann ich meine Genexpressionsdaten im zwei dimensionalen darstellen. Das ist der so genannte Hauptkomponentenplot oder **PCA-Plot**. Was kann ich aber darin sehen? Ich kann zum

Beispiel sehen, ob meine Samples **in Gruppen aufsplitten**. In der linken Skizze der Abbildung 7 habe ich 5 behandelte Zelllinien und 5 Kontrollgruppen. Diese 10 Punkte spiegeln die Genexpression von allen Genen wider, aber runtergebrochen auf den zwei dimensionalen Raum. Ich kann also schonmal mit dem Plot prüfen, ob die Genexpression sich überhaupt in Gruppen aufteilt.

Abbildung 7: PCA-Plot.



Es ist also ein erster Indikator, dass Einzelgene differentiell exprimiert werden zwischen den Gruppen. Dann kann ich den PCA-Plot auch für eine **Qualitätskontrolle** verwenden (Abbildung 7 mittlere Skizze). Nehmen wir eine Situation wie in der Mitte von Abbildung 7. Da fällt ein behandeltes Sample (roter Punkt) in die Region der Kontrollgruppen-Samples (schwarze Punkte). Umgekehrt haben wir ein Sample aus der Kontrollgruppe unter den behandelnden Samples. Dann würde ich zum Biologen gehen und fragen, ob er die Daten alle richtig gelabelt hat. Vielleicht hat er ein Kontroll-Sample mit einem behandelten Sample vertauscht. Es kann ja mal vorkommen, dass man Daten falsch beschriftet. Deshalb ist der PCA-Plot eine Qualitätskontrolle, die solche Fehler im Labor sichtbar machen kann. Ebenfalls kann ich auch *Ausreißer* entdecken (rechte Skizze von Abbildung 7). Wir sehen in der rechten Skizze der Abbildung 7 zum Beispiel unten links im Graphen einen roten Punkt, der weit entfernt von den anderen Punkten geclustert wurde. Jetzt kann man sich fragen, ob bei diesem Sample/ roten Punkt die Messung schief ging. Vielleicht habe ich auch Patientendaten und ich habe einen Patienten einer Gruppe zugeordnet, in die er vielleicht nicht wirklich reingehört. Vielleicht hat der Patient eine ganz andere Erkrankung und ich habe den Patienten in die Gruppe einer für Ihn falschen Erkrankung reingepackt. So können wir Ausreißer detektieren. Das nennt man eine **explorative Datenanalyse**. Damit kann man **die korrekte Gruppierung und Qualität überprüfen.**

s.18. **Zwei-Gruppen Versuchsdesign – Daten Exploration.**

Es gibt noch ein zweites Verfahren. Das Verfahren ist Ihnen vertraut, weil wir schon über phylogenetische Bäume gesprochen haben. Ich habe auch die Möglichkeit meine Genexpressionsprofile zu clustern. Gehen wir dazu zurück zur Abbildung 1. Genauso wie wir Sequenzen in einem hierarchischen Baum geclustert haben für die phylogenetische Analyse (da haben wir nach Ähnlichkeiten in den Sequenzen geschaut), können wir hier nach Ähnlichkeiten in den Spalten des Datensatzes (Abbildung 1) schauen. Also **wie ähnlich sind sich die Kontrollindividuen und die behandelten Individuen oder kranken Individuen untereinander**. Ich könnte zum Beispiel die Korrelation zwischen den Spalten oder die Distanz berechnen. Dann habe ich eine Distanzmatrix und basierend auf dieser Distanzmatrix kann ich einen Baum bilden, wie wir es kennengelernt haben bei der phylogenetischen Analyse (Neighbor-Joining). Neben dem Neighbor-

Joining benutzt man hier aber noch zwei weitere Algorithmen: complete linkage und single linkage. Ich werde gleich noch zweigen, wo der Unterschied ist.

Abbildung 8: hierarchische Clusteranalyse mit Bäumen.

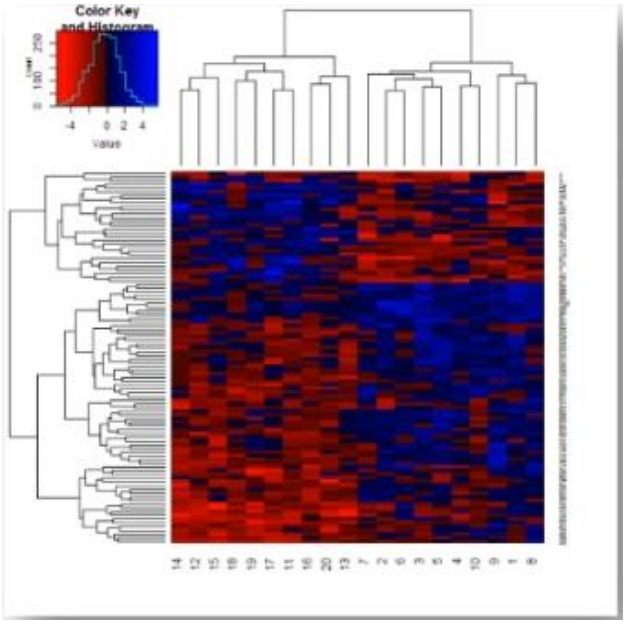
(Abbildung 8) Wenn ich mit complete linkage clustere, kann zum Beispiel sehen, ob sich meine Gruppen schön auftrennen. In der linken Skizze der Abbildung 8 sieht man zum Beispiel, dass sich die Kontroll-Samples alle links clustern (“3”). Die behandelten Samples clustern sich rechts im Baum (“2”). In der mittleren Skizze der Abbildung 8 haben wir genau wie bei der PCA ein Sample aus der Kontrollgruppe (“3”), der unter die behandelten Proben geclustert wurde. Das heißt, genau wie mit der PCA kann ich bei der hierarchischen Clusteranalyse eine Qualitätskontrolle betreiben. Die ersten beiden Skizzen in Abbildung 8 wurden mit Complete Linkage geclustert. Das ist ein bestimmtes Verfahren basierend auf der Distanzmatrix, mit der die Bäume gebaut werden. Wenn ich single Linkage verwende (rechte Skizze Abbildung 8), dann habe ich ein gutes Mittel, um Ausreißer zu erkennen. Alle drei Bäume in Abbildung 8 basieren auf den gleichen Daten, aber durch einen anderen Cluster-Logarithmus kriege ich unterschiedliche Baumstrukturen. Bei Single Linkage kann ich sehr gut sehen, dass wir ein Sample haben, der von den anderen Samples abweicht. Auch hier kann man hinterfragen, ob es eine Fehlmessung ist oder ob der Patient eine andere Erkrankung hat.

S.19. **Zwei-Gruppen Versuchsdesign – Daten Exploration.**

Man kann sich die Tabelle auch in Form einer Hitzekarte oder “**Cluster Heatmap**” (Abbildung 9) visualisieren lassen. Typischerweise benutzt man eine geclusterte Heatmap. Das heißt, die Clustert Bäume aus Abbildung 8 nutze ich auch, um die Spalten im Datensatz (Abbildung 1) neu anzuordnen. Man kann auch die Clusteralgorithmen (Neighbopr-Joining, complete und single linkage) nutzen, um Gene mit ähnlichen Genexpressionsprofilen zu clustern. Diese Gene werden dann an den gleichen Ast gehangen. Ich sortiere also die Zeilen und Spalten aus Abbildung 1 also um, damit es zu den Cluster-bäumen aus Abbildung 8 passt. Durch eine Farbcodierung kann ich dann zeigen, wie stark ein Gen exprimiert wird. Sehr helles blau in Abbildung 9 heißt, dass das Gen stark exprimiert wird, während rot heißt, dass das Gen kaum exprimiert wird. Dann kann ich in der CLuster Heatmap im Idealfall bestimmte Blöcke erkennen. Ich habe zum Beispiel in Abbildung 9 eine Region oben links, wo wir sehr viel blau haben und das bedeutet, dass in diesen Samples ein Block an Genen stark exprimiert wurde. Rechts Oben haben wir direkt daneben eine sehr rote Region/Block. Sagen wir mal, das waren die behandelten Samples. Dort waren die Gene sehr schwach exprimiert. Man muss aber dazu sagen, dass die Cluster Heatmaps nicht so Sinnvoll sind, um sie auf die gesamte Datenmatrix anzuwenden. Bei 40.000 Genen wäre es sehr schwer eine Distanzmatrix zu erstellen, die alle Gene abdeckt. Von der Effizienz her ist es schon fast unmöglich. Der Computer kann nicht jedes Gen mit jedem Gen korrelieren. Die Distanzmatrix ist kein gutes Maß für die Korrelation

zwischen den Genen. Typischerweise nutzt man deshalb Heatmaps für selektierte Gene. Man schaut sich damit also nur bestimmte Pathways an oder Gene, die in dieselbe Gruppe gehören. Alle Gene können wir damit nicht darstellen.

Abbildung 9: Heatmap (X-Achse Samples, Y-Achse Merkmale/Gene).



Jetzt haben wir Datenvorverarbeitung und Daten Exploration durch PCA und Clusteranalyse behandelt. Jetzt sind wir endlich so weit, dass wir zur eigentlichen Analyse kommen. Wir reden jetzt über die Selektion von Genen, die zwischen den Gruppen unterschiedliche Expressionen aufweisen und damit differenziell exprimiert werden.

s.20. Gen-weise Analyse.

Abbildung 10: Berechnung des Fold Changes.

- **Fold Change**

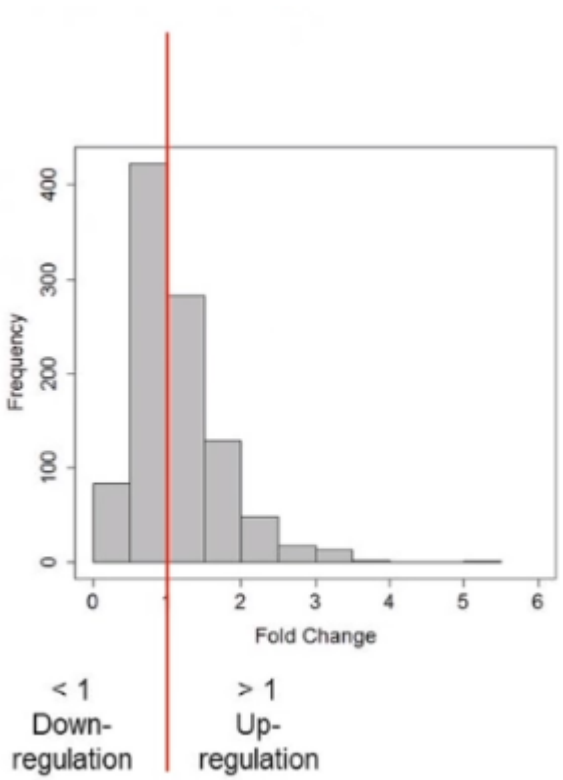
$$\frac{\text{Mittelwert}_{\text{Erkrankte}}(35, 45, 53, 34) = 41.75}{\text{Mittelwert}_{\text{Kontrolle}}(24, 16, 12, 17) \quad = 17.25} = 2.42$$

Wir wechseln jetzt auf die Ebene einzelner Gene. Wir wollen wissen, welche Gene hoch oder runterreguliert sind durch einen externen Effekt wie Infektion oder eine Erkrankung. Wir brauchen für jedes Gen ein Maß, welches **die Stärke der Expressionsänderung** angibt. Das übliche Maß, das man dafür verwenden, ist das so genannte **Fold Change**. Um das wie viel fache ändert sich etwas? Das gibt der Fold Change an. Der Fold Change lässt sich einfach berechnen (Abbildung 10 für Rechnung). Sagen wir mal, wir haben 4 Samples bei der Erkrankten und 4 Samples bei der Kontrollgruppe. Ich rechne dann jeweils den Mittelwert aus und dividiere beide Gruppen durcheinander. In Abbildung 10 bekomme ich dann ein Fold Change von 2.42. Die Interpretation wäre wie folgt: Bei den Erkrankten ist dieses Gen um das 2.42-fache erhöht und wird damit um das

2.42-fache höher exprimiert im Vergleich zur Kontrollgruppe. **Der Fold-Change gibt also das x-Fache einer erhöhten oder erniedrigten Genexpression im Vergleich zur Kontrollgruppe an**. Wenn wir das jetzt umdrehen und von dieser Rechnung den Kehrwert nehmen, dann habe ich einen Fold Change von 1/ 2.42 oder ½, um es einfacher zu machen. Das heißt, eine 2-fache Hochregulation oder ein Fold Change von 2 entspricht einem Fold Change von 0,5 (½) bei einer entsprechenden Runterregulierung.

s.21. Gen-weise Analyse

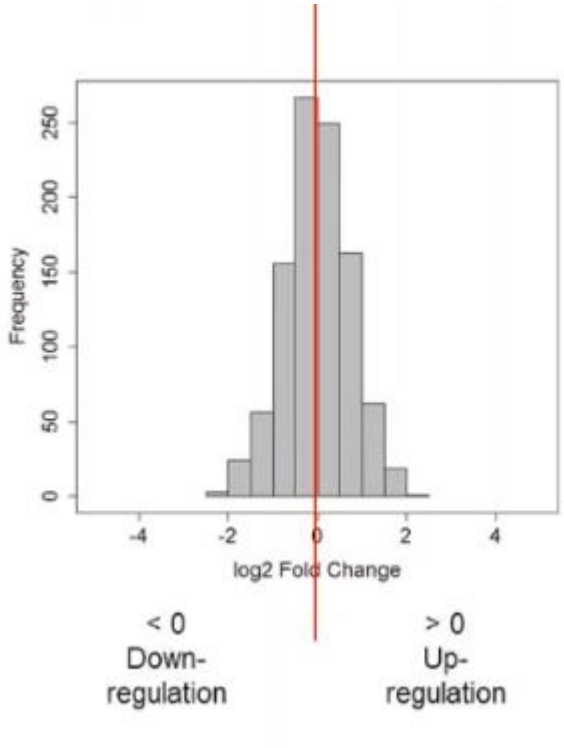
Abbildung 11. Gen-weise Analyse: Fold Change Histogramm.



Wir sehen genau das nochmal dargestellt in Abbildung 11. Wir haben hier das Histogramm der Fold Changes für alle 40.000 Gene in einem Experiment. Gene, die hochreguliert sind, nehmen den Bereich von 1 bis Unendlich ein (also alles rechts von der roten Linie in Abbildung 11). Das liegt daran, dass ein Quotient von positiven Werten theoretisch unendlich groß sein kann (Zähler ist Größer als der Nenner und je größer der Zähler, desto größer ist das Ergebnis). Wenn ich etwas sehr, sehr Großes durch etwas sehr, sehr kleines teile, dann wird der daraus resultierende Quotient sehr groß sein. Ein Fold Change von **1** heißt, dass ich dasselbe Expressionsverhältnis zwischen Kranken und Kontrollgruppe habe. Das bedeutet, dass ich in beiden Gruppen die gleiche Expression habe. Bei herunter regulierten Genen (Down Regulation) können die Fold Changes nur einen Wert **zwischen 0 und 1** annehmen. Das liegt daran, dass ich für Sie den Kehrwert des Fold Changes nehme. Wenn ich den Kehrwert eines Quotienten von positiven Werten nehme, dann kann das Ergebnis nur zwischen 0 und 1 liegen. Wenn ich bei einer 5-fachen Hochregulation den Kehrwert nehme, dann habe ich 1/5. Also sieht man schon daran, dass der Wert nie unter 0 und nie über 1 steigen kann. Das ist aber für das Ablesen der Ergebnisse etwas ungünstig. Es wäre schöner, wenn Hoch- und Runter-regulierte Gene symmetrisch wären. Wie kann ich eine Symmetrie in Abbildung 11 erreichen? Das erreiche ich hier wieder durch eine Logarithmierung.

s.22. Gen-weise Analyse

Abbildung 12: Fold Changes nach Logarithmierung



Wenn ich die Fold Changes logarithmiere (man spricht dabei von " log Fold Change"), dann habe ich eine symmetrische Verteilung (Abbildung 12). Logarithmus von 1 ist 0. Das heißt, Gene, die sich zwischen Erkrankten und der Kontrollgruppe nicht unterscheiden, haben ein log Fold Change von 0. Gene, die hoch reguliert sind, haben einen positiven Log Fold Change. Gene, die runter reguliert werden, haben einen negativen Log Fold Change. Ein Log Fold Change von +2 korrespondiert sehr schön mit einem log Fold Change von – 2 bei einer Runterregulierung. Das macht das Ablesen sehr einfach. Ich kann schon am Vorzeichen des log Fold Changes erkennen, ob ein Gen hoch reguliert wird oder runterreguliert wird. Typischerweise verwendet man den log2 –also Logarithmus zur Basis 2. Das hat noch ein paar gute Eigenschaften für die weiteren Berechnungen, die wir im Verlauf der Analyse noch kennenlernen werden.

S.23-24. Gen-weise Analyse.

Ich berechne also diesen Fold Change oder besser gesagt den Log Fold Change für jedes Gen, aber nicht nur das: **Der log Fold Change zeigt mir nur an, um das wie vielfache ein Gen seine Expression verändert**. In den Fold Change fließen auch **nur die mittleren Genexpressionen** ein. Ich kriege keine Informationen über die Varianz der Gene. Um das zu berücksichtigen, berechnet man nicht nur den Log Fold Change, sondern man berechnet auch einen **statistischen Test**. Am Ende eines statistischen Tests kriege ich einen **P-Wert** heraus. Dieser sagt mir, wie wahrscheinlich es ist, dass **kein Effekt** vorliegt. Ein sehr kleiner P-Wert heißt, dass wahrscheinlich ein Effekt vorliegt bzw. Die Wahrscheinlichkeit, dass kein Effekt vorliegt, ist sehr gering. Also, pro Gen berechne ich einen statistischen Test (siehe Abbildung 13), der mir die Expressionswerte in der Kontrollgruppe und bei den Erkrankten vergleicht, und ich quantifiziere die Expressionsänderungen über den Log Fold Change. Der Vorteil vom P-Wert ist, dass er anders als der Log Fold Change nicht nur die Mittelwerte

einbezieht, sondern er bezieht auch Information über Varianz und Stichprobenumfänge mit ein. Wenn ich einen hohen Stichprobenumfang habe, dann besteht auch eine höhere Wahrscheinlichkeit, dass ich einen kleinen Effekt erkenne. Der Nachteil vom P-Test ist, dass man keine Quantifizierung hat. Dafür braucht man dann den Log Fold Change.

Abbildung 13: Abbildung 1 nach Einbezug des P-Tests und Log Fold Changes.

Gen	Kontrollgruppe			Erkrankte			p	log Fold Change
	1	...	n_1	1	...	n_2		
1	7.6	...	7.5	4.2	...	5.1	$2.7 * 10^{-10}$	-1.35
2	5.0	...	4.9	8.2	...	7.3	$4.7 * 10^{-09}$	-1.11
3	3.9	...	4.2	8.2	...	7.5	$9.4 * 10^{-09}$	+1.06
4	5.9	...	6.2	1.8	...	1.0	$1.1 * 10^{-08}$	-0.97
5	5.9	...	5.9	1.3	...	2.3	$1.9 * 10^{-09}$	-0.89
...	...	...	...	...	...	...	...	...
d	9.4	...	9.3	8.7	...	9.2	$9.9 * 10^{-01}$	0.01

Hat jemand eine Idee, wenn wir beim Microarray metrisch kontinuierliche Werte haben, welchen statistischen Test man anwenden könnte pro Gen? Welchen Test könnte ich verwenden, um zu sehen, ob es einen Unterschied gibt zwischen Erkrankten und Kotrollgruppe bei Gen 1 zum Beispiel? Es wird eine bestimmte Form des **T-Tests** angewendet. Wenn wir Count Daten haben, weil wir RNA-Seq verwendet haben, dann können wir den T-Test nicht mehr nutzen, weil der T-Test eine Normalverteilung voraussetzt. Bei Count-Daten nutzt man eine bestimmte Form des **Binomialtests** an. Count-Daten lassen sich nämlich als negative Binomialverteilung oder Poisson-Verteilung modellieren.

S.25 Gen-weise Analyse.

Abbildung 14: Standard T-Test.

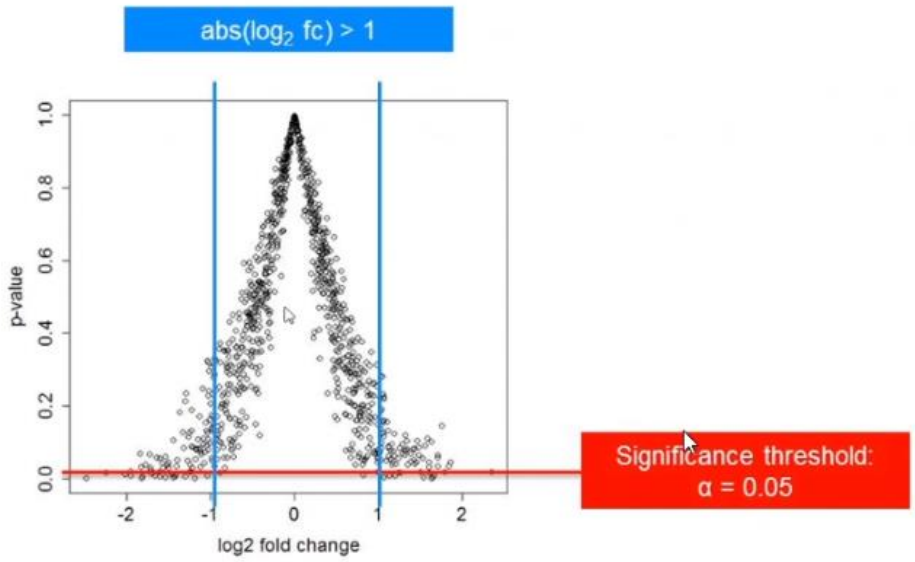
$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_{X_1X_2} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Hier (Abbildung 14) haben wir den klassischen T-Test. Es wird aber nicht genau dieser T-Test verwendet, sondern eine Variante. Denn das Problem ist, dass wir die Mittelwerts-Differenz im Zähler (Abbildung 14) - also Mittelwert von Gruppe 1 Minus Mittelwert von Gruppe 2- haben. Im Nenner haben wir eine gepolte Varianz (s_{x1x2}) aus beiden Gruppen und in der Wurzel haben die Stichprobenumfänge beider Gruppen. Wenn ich jetzt Tausende von Genen gleichzeitig teste (Multiple Testsituation), dann besteht die Gefahr, dass bei manchen Genen sehr kleine Varianzen auftreten, wodurch die Test-Statistik unglaublich groß wird. Das ist einfach ein Zufallseffekt, der auftritt, wenn man Tausende Gene testet. Das erhöht insgesamt die Gefahr, dass ich falsch positive Ergebnisse bekomme. Dann schlussfolgere ich, dass es einen Unterschied in der Geneexpression bei einem Gen gibt, obwohl es nicht wahr ist. Die Statistiken, die man für Genanalysen verwendet, sind die **moderierten T-Statistiken.** Hier wird die Varianz s_{x1x2} geschrumpft. Damit kann man vermeiden, dass per Zufall sehr große Test Statistiken entstehen. So vermeidet man dann auch falsch positive

Ergebnisse. Wir werden uns in der Software-Übung ein Package anschauen, welches diesen T-Test für uns übernimmt.

S.26. **Gen-weise Analyse**

Abbildung 15:Volcano Plot



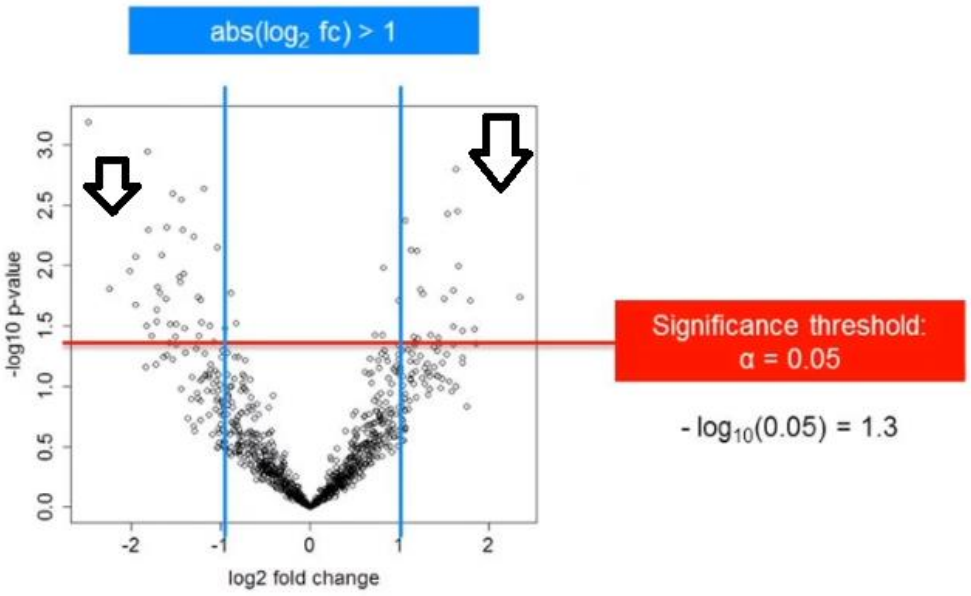
Wenn ich dann für jedes Gen den P-Wert und den log Fold Change Wert berechnet habe, dann kann ich mir das Ganze nochmal visualisieren lassen (Abbildung 15). Zur Visualisierung haben wir den so genannten **Volcano Plot**. Auf der X-Achse habe ich den Log2 Fold Change. Auf der Y-Achse sind die P-Werte. Jeder Punkt in Abbildung 15 repräsentiert ein Gen. Beim P-Wert – das kennen Sie bereits aus dem statistischen Test – hat man einen **Schwellenwert** für die statistische Signifikanz. Als Konvention verwendet man **Alpha =0,05**. Das heißt, alle Gene, die einen P-Wert kleiner als 0,05 haben, würde man als Signifikanz differentiell exprimiert zwischen den zwei Gruppen betrachten. Für den Log Fold Change kann ich auch einen Schwellenwert einführen. Ich habe in der Abbildung +/- 1 als Threshold verwendet. Dann würde man sagen, dass Gene, die einen Threshold über 1 und unter 1 haben, für uns interessant sind. Das heißt alle Gene, die rechts von der blauen +1 Linie liegen, sind für uns interessant, und Gene, die links von der blauen –1 Linie liegen, sind für uns interessant. Uns interessieren diese Gene, da wir je nur die Gene beobachten wollen, die entweder sehr stark hochreguliert wurden durch die Erkrankung oder sehr stark runterreguliert wurden durch die Erkrankung. Der Nachteil dabei ist, dass wir ja auch die Signifikanz mit einbeziehen und damit wollen wir nur die Gene haben, die rechts und links von den blauen Log Fold Change Linien liegen und dabei noch unter der roten Linie liegen. Das bedeutet aber, dass die Rechtecke, für dich mich am meisten interessiere (rechts unten und links unten), relativ klein sind. Ich möchte da jetzt reinzoomen, damit wir zum Beispiel die Gene beschriften können. Dazu verwende ich auch hier wieder den Logarithmus.

S.27.

Ich logarithmiere den P-Wert und zwar mit dem 10er Logarithmus und multipliziere mal Minus 1 (Abbildung 16). Dadurch dreht sich die ganze Abbildung um. So sind die Gene, die über der roten Linie liegen signifikant. Für mich interessant sind jetzt die Rechtecke rechts und links oben (siehe Pfeile). Das ist also eine ziemlich gute Möglichkeit, um in die kleinen Rechtecke aus Abbildung 15 rein zu zoomen. So sieht man dann die signifikanten Gene dann besser und kann sie besser

beschriften oder anders mit ihnen interagieren und arbeiten. Das ist im Grund nur eine Visualisierungsmethode, aber man kann auch ganz gut sehen, wie viele Gene wirklich signifikant sind und wie stark die Expressionsänderung wirklich ist pro Gen.

Abbildung 16: Volcano Plot nach Logarithmierung und *-1

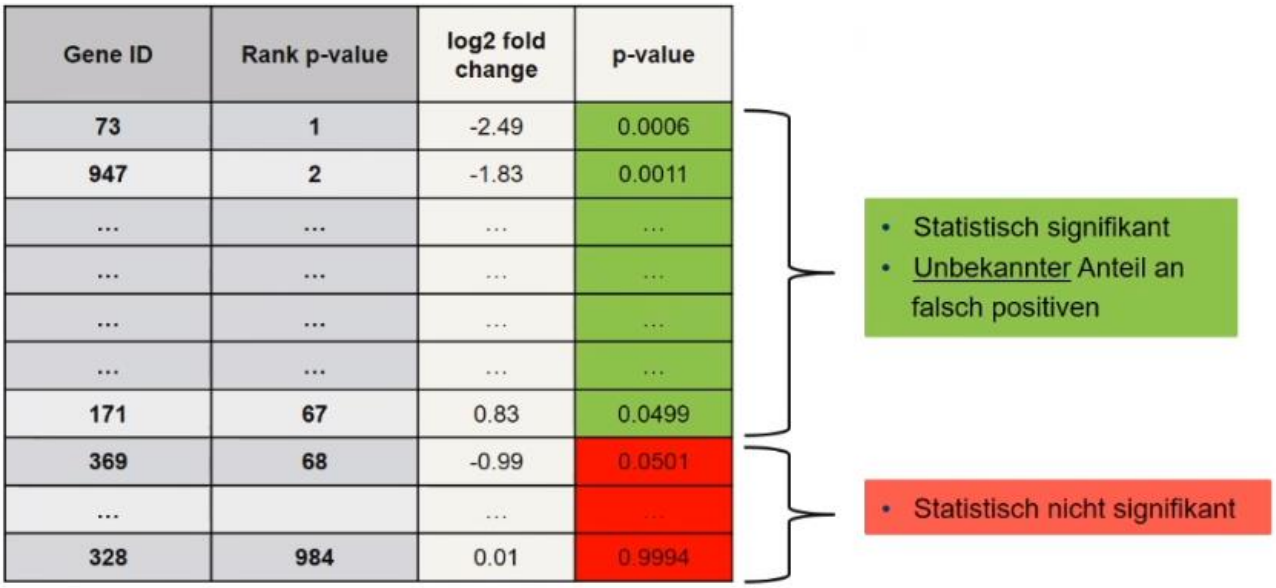


s.28-31. Gen-weise Analyse.

Jetzt bringt das statistische Testen, wenn ich tausende Tests durchführe, noch ein weiteres Problem mit sich. Ich habe ja bereits den Begriff Multiples Hypothesentesten gebracht. Eigentlich möchte ein Statistiker nur ganz wenige Hypothesen testen. Denn bei jedem statistischen Test habe ich die Gefahr, dass ich zu einer falschen Schlussfolgerung komme. Hier noch mal kurz der Einwurf: Wie generieren wir eigentlich in der Wissenschaft Erkenntnisse basierend auf Daten? Wir analysieren die Daten und ziehen Schlussfolgerungen daraus. Dabei kann es aber passieren, dass ich als Wissenschaftler auch falsche Schlussfolgerungen ziehe, weil ich zum Beispiel Streuungen in meinen Daten habe und ich generiere Daten aus einer Stichprobe, die vielleicht nicht wirklich repräsentativ ist. Experimentelle Daten können also auch zu einem Trugschluss führen. Passen Sie also auf, dass die Daten Sie nicht in die falsche Richtung leiten. Beim Hypothesentesten kann es eben sein, dass ich basierend auf meiner Analyse auf einen Effekt schließe, aber der Effekt ist in Wirklichkeit gar nicht da. Das wäre ein falsch positives Ergebnis. Alternativ kann es sein, dass meine Daten nicht ausreichen, um einen Effekt zu erkennen. Dann hätte ich ein Falsch Negatives Ergebnis. Beim Multiplen Testen habe ich insgesamt eine sehr hohe Wahrscheinlichkeit, dass ich viele falsch positive Ergebnisse habe. Nehmen wir mal an, ich bekomme 500 Gene heraus, wo mir der P-Wert sagt, dass sich die Genexpression zwischen Erkrankten und Gesunden signifikant geändert hat. Unter diesen 500 Genen wird es aber viele falsch positive Ergebnisse geben, weil ich bei jedem statistischen Einzeltest die Gefahr habe, ein falsch positives Ergebnis zu produzieren.

Ich habe die Gene in Abbildung 17 nach dem P-Wert sortiert. Im grünen Bereich fängt es mit dem kleinsten P-Wert an und das Grüne Feld hört bei einem P-Wert auf, der knapp unter der 0,05 Schwelle ist. Das wären meine Werte, die signifikant sind. Die Roten sind nicht signifikant. Unter den Signifikanten (Grün) habe ich einen unbekannten Anteil an Falsch-Positiven.

Abbildung 17: Signifikante und Nicht Signifikante Werte in einer Tabelle.



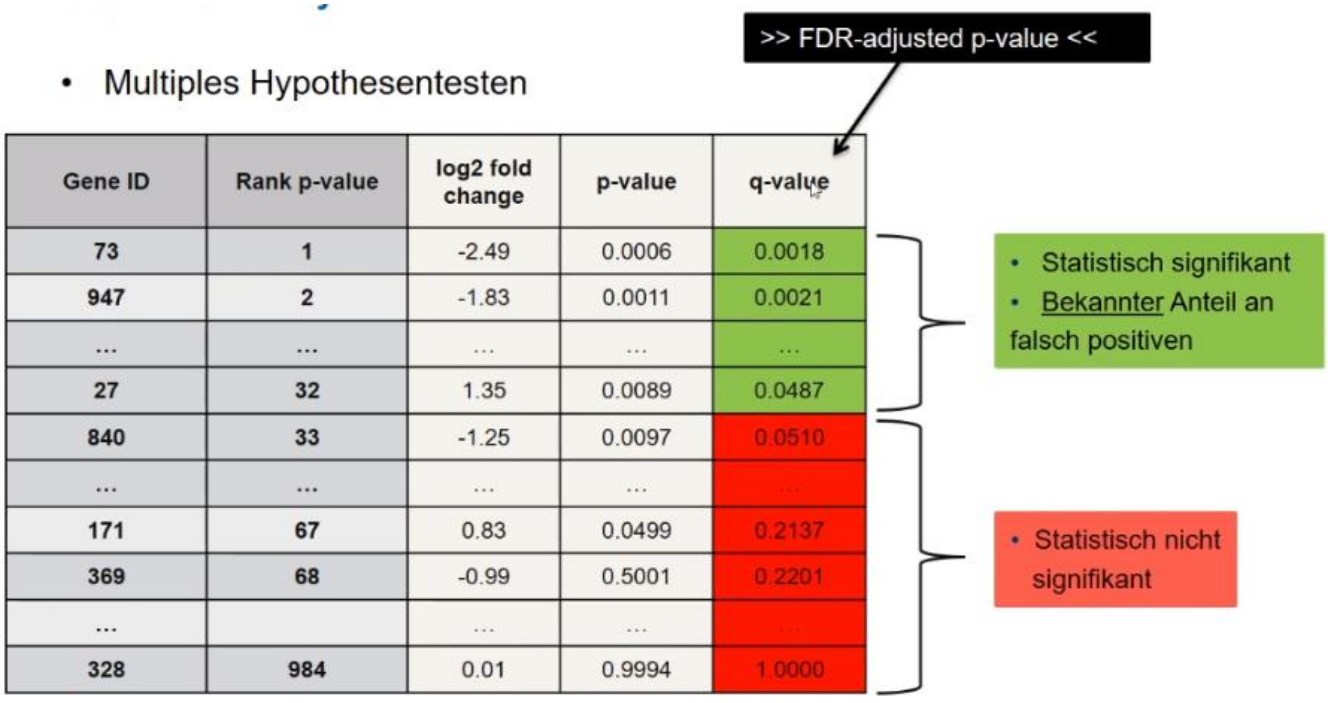
Gene ID	Rank p-value	log2 fold change	p-value
73	1	-2.49	0.0006
947	2	-1.83	0.0011
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
171	67	0.83	0.0499
369	68	-0.99	0.0501
...		...	...
328	984	0.01	0.9994

Wie kann ich jetzt aber diesen Anteil kontrollieren?

s.32.

Es gibt einen so genannten **korrigierten P-Wert**. Man spricht von **FDR-adjustierten P-Werten.** FDR steht für “false discovery rate”. Also, ich möchte die Anzahl an Falsch-Positiven reduzieren oder zumindest kontrollieren. In mancher Literatur wird der adjustierte P-Wert auch als **Q-Wert** bezeichnet (Abbildung 18).

Abbildung 18: Tabelle aus Abbildung 17 mit Q-Wert.



Gene ID	Rank p-value	log2 fold change	p-value	q-value
73	1	-2.49	0.0006	0.0018
947	2	-1.83	0.0011	0.0021
...	...	...	...	...
27	32	1.35	0.0089	0.0487
840	33	-1.25	0.0097	0.0510
...	...	...	...	...
171	67	0.83	0.0499	0.2137
369	68	-0.99	0.5001	0.2201
...		...	...	...
328	984	0.01	0.9994	1.0000

In biologischer und Medizinischer Literatur werden Sie Q-Werte kaum finden. Dort werden Sie “FDR-adjusted p-values" genannt. Dieser adjustierter P-Wert ist in der Regel größer als der “rohe” P-Wert.

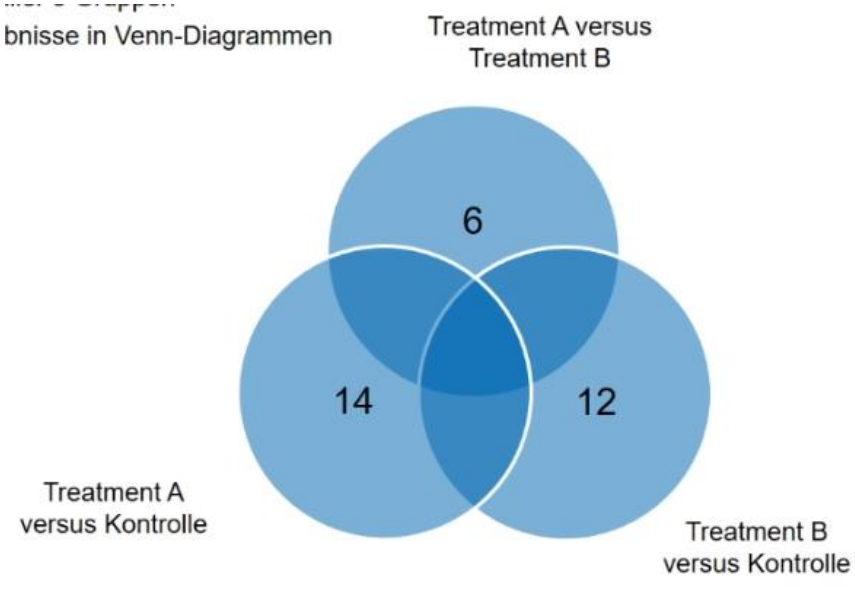
Ich erhöhe also meine P-Werte. Das heißt, am Ende sind weniger Gene signifikant. Der einzige Vorteil, der am Ende in der Adjustierung liegt, ist, dass ich Falsch-Positive reduziere und ich kann den Anteil an Falsch-Positiven kontrollieren. Dieser Adjustierungsalgorithmus ist so ausgelegt von der Mathematik her, dass ich angeben kann, dass ich eine FDR von 5 % haben möchte und am Ende werde ich unter den Signifikanten (grün) 5 % an Falsch-Positiven haben. Es bleiben also 5 % Falsch-Positive drinnen, aber ich weiß zumindest wie hoch der Anteil an Falsch-Positiven ist. Bei den rohen P-Werten habe ich hingegen einen hohen Anteil an Falsch-Positiven und dazu noch weiß ich nicht, wie viele davon Falsch-positiv sind. Nach der Adjustierung habe ich einen kleinen Anteil und ich weiß, wie groß der Anteil an Falsch-Positiven ist. Damit können wir also falsche Schlussfolgerungen reduzieren. Ich möchte hier anmerken, das, was wir in der Wissenschaft schlussfolgern, basiert nicht nur ausschließlich auf den Daten. Sie haben in jeder Analyse und in jedem Experiment Annahmen. Man betrachtet Daten nie ohne Annahmen. Ich nehme zum Beispiel an, dass meine Datensätze einer Normalverteilung folgen, auch wenn ich es zu dem Punkt noch nicht beweisen kann. Annahmen und Daten führen zu wissenschaftlichen Schlussfolgerungen (Take Home Botschaft).

s.33. Gen-weise Analyse

Bisher sind wir davon ausgegangen, dass wir ein Zwei-Gruppen Experiment haben. Im Labor macht man aber auch mal Experimente, bei denen man mehr als Zwei Versuchsgruppen hat. Man kann zum Beispiel eine Kontrollgruppe haben und dann zwei verschiedene Behandlungen. Dann habe ich 3 Gruppen. Auch wenn ich reale Patientendaten habe, kann es sein, dass ich eine Kontrollgruppe gegen drei verschiedene Medikamente vergleiche. Wie geht man dann bei der Datenauswertung vor? Theoretisch könnte ich für jedes Gen eine Varianzanalyse machen, die mir sagt, ob es zwischen den Gruppen einen Unterschied gibt. Üblicherweise geht man aber so vor, dass man bei mehr als zwei Gruppen jede Gruppe mit jeder Gruppe vergleicht. Dabei nutzt man die Methoden, die wir in dieser Vl kennengelernt haben. Auch hier haben wir einen P-Wert, einen adjustierten P-Wert und Fold Changes. Dann vergleiche ich Behandlung 1 gegen Kontrolle, dann Behandlung 2 gegen Kontrolle und dann vergleiche ich Behandlung 1 und 2 untereinander. Das heißt, bei drei Gruppen kriege ich drei Listen an Genen heraus, die mir sagen, welche Gene zwischen den drei Gruppen differentiell exprimiert werden.

s.34. Gen-weise Analyse.

Abbildung 19: Venn-Diagramm bei 3 Gruppen



Hinterher kann ich das Gante in einem **Venn-Diagramm** darstellen (siehe Abbildung 19). Ich habe also eine gewisse Menge an Genen, die in allen 3 Vergleichen signifikant exprimiert werden (Schnittstelle von allen Kreisen in Abbildung 19). Dann habe ich zum Beispiel nur 12 Gene, die nur beim Vergleich von Behandlung B gegen die Kontrolle signifikant sind (keine Schnittstelle mit den anderen Kreisen in Abbildung 19). Also, anstatt mit Varianzanalytischen Methoden komplexe Designs auszuwerten, macht man ganz Schlicht **Paarvergleiche** und man gibt dann im Venn-Diagramm an, wo es Schnittmengen zwischen den Vergleichen gibt.

s.35.

Abbildung 20: Mengenvergleich bei mehr als 3 Proben (oben 4 Gruppen

Wenn ich 4 Vergleiche mache (Abbildung 20, obere Skizze), dann gibt es auch noch solche Mengendiagramme. Bei 5 Vergleichen ist es auch noch möglich Mengendiagramme anzugeben (unten, Abbildung 20). Ab einer bestimmten Zahl hat man aber so viele Teilschnittmengen, dass eine genaue Interpretation sehr schwer wird. Hinter jeder Schnittmenge steht eine bestimmte Menge an Genen und ab einer bestimmten Gruppenmenge und Vergleichsmenge kann man kaum sagen, welches Gen in welchem Vergleich signifikant differenziert exprimiert wurde. Hier ist mein Ratschlag: Machen Sie lieber mehrere kleine Experimente, die nicht so komplex sind. Je komplexer das Experiment ist, desto mehr Faktoren müssen berücksichtigt werden. Das macht ein Experiment unsicherer. Ein kleines Experiment ist sicherer und liefert Ergebnisse, die dich leichter validieren lassen.

S.36. Datenbanken.

Wir kennen bereits die Sequenzdatenbanken. Für Experimente mit Genexpressionsdaten gibt es auch Datenbanken. Die wichtigsten sind **ArrayExpress** und **Gene Expression Omnibus**. Dort liegen experimentelle Daten, die man sich frei herunterladen kann. Wenn ich ein Experiment publiziere als Biologe oder Mediziner, dann verlangt das Journal von mir, dass ich die Rohdaten oder prozessierten Daten in einer öffentlichen Datenbank hochlade, damit andere mein Experiment nachrechnen und überprüfen können. Sie müssen also alle Daten hinterlegen, wenn Sie etwas publizieren wollen.

Mal angenommen, ich habe 40.000 Gene für ein Experiment genutzt. Dann habe ich 200 Hundert Gene, wo ich sagen kann, dass sie sich durch die Erkrankung signifikant verändert haben in ihrer Expression. Was mache ich jetzt mit diesen signifikanten Genen als Biologe? Man schaut nach, ob Sie funktionell zusammenhängen. Vielleicht hängen Sie alle im selben Stoffwechsel zusammen. Die 200 Gene, die ich habe, sind in einer Reihenfolge sortiert. Ich schaue mir zuerst die Top-Gene an. Die Gene mit den kleinsten P-Werten schaue ich mir zuerst an. Dann gehe ich auf Google Scholar. Google mein Gen und schaue nach, ob es nicht bereits Paper zu dem Gen gibt, die mir weiterhelfen können. Bei der funktionellen Einteilung der Gene muss ich die Gene annotieren. Dann schaue ich mir an, wie die Gene mit bestimmten Pathways zusammenhängen. Diese sogenannte **Genset-Analyse** ist das Thema der nächsten VL. Da schauen wir uns an, ob es Gensets/Pathways gibt, die innerhalb der differenziell exprimierten Gene besonders hervortreten.