Softwarewerkzeuge der Bioinformatik

Prof. Dr. Volkhard Helms
PD Dr. Michael Hutter, Markus Hollander,
Andreas Denger, Marcial Josef Paszkiel

Saarland University
Department of Computational Biology

Winter semester 2022/2023

Tutorial 9

12. Januar 2023

Gen-Annotation und Jupyter Notebooks

Exercise 9.1: Gene Set Enrichment Analysis

In dieser Übung werden Sie eine Pathway Enrichment Analysis auf einem Datensatz ausführen der Microarray Genexpressionsdaten von Adulter T-Zell-Leukämie (ATL) Patienten enthält, sowie Referenz-Samples von Patienten ohne ATL. Pathway annotationen beschreiben die molekularen Interaktionen die ein Gen bzw. dessen Produkt ausführt, in und um die Zelle. Das Ziel dieser Art von Analyse ist die Krebs/nicht-Krebs samples zu vergleichen, und nach Pathways zu suchen die signifikant angereichert oder verringert sind in Krebszellen, im Vergleich zu gesunden Zellen. Das kann uns ein besseres Verständnis der zellulären Mechanismen in Krebszellen geben, und aufzeigen wie sie sich von gesunden Zellen unterscheiden.

(a) Vorbereitung

(1) Gene Expression Omnibus (GEO) ist eine Datenbank auf der Genexpressions-Datensätze öffentlich zugänglich sind. Der ATL Datensatz den wir für diese Übung verwenden werden hat den GEO Identifier **GSE14317**. Klicken sie auf den Link und schauen sie sich den Eintrag in der Datenbank an. Wie viele ATL und Control-Samples wurden analysiert in dieser Studie? Welchem Gen haben die Forscher besondere Aufmerksamkeit geschenkt?

(2) Für diese Pathway-Analyse werden wir den Webservice *Genetrail* benutzen. Besuchen sie die Website unter https://genetrail.bioinf.uni-sb.de/ und starten sie eine *Transcriptomics* Analyse.

(3) Als nächstes müssen wir den Datensatz hochladen. Genetrail erlaubt uns einen GEO Identifier einzugeben, und lädt die Daten von dort automatisch herunter. Wählen sie *Enter a GSE File* und schreiben sie den Identifier in das Feld. Starten sie anschließend die Analyse.

(4) Im nächsten Fenster müssen die Samples ihren Gruppen zugeteilt werden. Die Samples deren Name mit *ATL* anfängt werden in die *Samples* Gruppe eingeteilt, die *CD* Samples kommen in die *Reference* gruppe. *Hinweis: Man kann mehrere Samples auswählen indem man Shift auf der Tastatur gedrückt hält.*

(b) Statistische Tests

(1) Die *Identifier-level statistic* wird benutzt um Gene zu finden die eine signifikante differenzielle Expression zwischen Krebszellen und gesunden Zellen zeigen. Wählen sie *Independent Students t-Test.*

(2) Die *Set-level statistic* berechnet ob die Gene die mit einem bestimmten Pathway assoziiert sind in den Daten angereichert sind. Hier nehmen wir eine Gene Set Enrichment Analysis (GSEA) in Form eines Kolmogorov-Smirnov-Tests. Dieser Test ist nicht parametrisch, was bedeutet dass er weniger Annahmen über die statistische Verteilung der Daten macht und damit robuster ist.

(3) Wählen sie nur *KEGG* und *Reactome* als Kategorien, scrollen sie nach oben und drücken sie *Start analysis.*

(c) Analyse der Ergebnisse

(1) Öffnen sie die Darstellung der Resultate indem sie auf *View* klicken.

(2) Laut der Publikation hatten die Gene **PCNA** und **BIRC5** eine hohe differenzielle Expression zwischen den zwei Gruppen. Wie vielen Reactome bzw. KEGG pathways gehören sie an? *Hinweis: Schreiben sie das jeweilige Gen-Symbol in die Suchleiste die auf der rechten Seite erscheint wenn man auf KEGG - Pathways oder Reactome - Pathways klickt.*

(3) Suchen sie nach Reactome pathways die **BIRC5** enthalten. Klicken sie auf *More...* neben dem Pathway mit dem Namen *SUMOylation of DNA replication proteins*, und suchen sie dort erneut nach **BIRC5**. Ein Klick auf den Eintrag bringt sie zu dem *genecards.com* Eintrag für dieses Gen. Warum könnte eine Hochregulierung dieses Gens etwas mit Krebs zu tun haben? Was ist der Name des Gens?

Exercise 9.2: Pathway & 3D Struktur Analyse

In diesem zweiten Teil der Übung werden wir uns einen der Pathways näher anschauen, und anschließend die 3D-Struktur eines Proteinkomplexes betrachten der Teil dieses Pathways ist.

(a) Rufen sie reactome.org auf und suchen sie nach *SUMOylation of DNA replication proteins.* Wählen sie das erste Suchergebnis. Klicken sie auf den Namen des Pathways in *Locations in the PathwayBrowser.*

(b) Wie wir in Übung 9.1 herausgefunden haben, ist das Gen **BIRC5** mit diesem Pathway annotiert. Nun müssen wir herausfinden wo in dem Pathway es sich befindet. Oben links finden sie ein Suchfenster. Suchen sie nach **BIRC5**, und wählen sie die *human* Variante die sich im *Nucleoplasm* befindet. Anschließend wird sich darunter ein kleines Fenster namens *Details* öffnen.

(1) Welchem Protein-Komplex gehört **BIRC5** an? Welche anderen Proteine sind in diesem Komplex?

(2) Wählen sie den Komplex aus, und klicken sie auf *Expression.* Benennen sie drei Gewebetypen in denen **BIRC5** eine hohe Genexpression hat.

(3) An welcher Reaktion nimmt **BIRC5** hier Teil? Betrachten sie die *Inputs*, *Outputs* und *Catalysts* der Reaktion und beschreiben sie was mit den Molekülen passiert. *Hinweis: Klicken sie auf die Reaktions-Pfeile der Reaktion.*

(c) Ein weiterer Pathway dem **BIRC5** angehört enthält Proteine die von **TP53** gehemmt werden. Wir können den Webservice *STRING-db* benutzen um mehr Informationen über dieses Protein herauszufinden.

(1) Rufen sie https://string-db.org/ auf und wählen sie *Search.* Schreiben sie das Gensymbol **TP53** in das Feld, und wählen sie *Homo sapiens* als den Organismus.

(2) Klicken sie auf *Settings.* Wählen sie nur *Experiments* und *Databases* als *active interaction sources*, setzen sie die *minimum required interaction score* auf *high confidence*, und drücken sie anschließend *Update.*

(3) Klicken sie auf den Knoten mit dem Namen **TP53** um eine kurze Zusammenfassung zu bekommen. Welche Rolle spielt dieses Gen in vielen Tumor-Arten und was ist seine Funktion?

(4) Fügen sie dem Netzwerk um **TP53** weitere Proteine hinzu indem sie zwei mal auf *+More* klicken. Danach sollten insgesamt 20 Proteine angezeigt werden.

(5) Wechseln sie zu dem *Analysis*-Tab. Hier können sie Genannotationen finden die in diesem Teilnetzwerk von 20 Proteinen angereichert sind. Sie können auf eine Annotation klicken um die Proteine die damit annotiert sind oben im Netzwerk einzufärben. Welche der 20 Proteine sind assoziiert mit *small cell lung cancer*? *Hinweis: Erweitern sie die Liste der KEGG-Pathways.*

Exercise 9.3: Jupyter Notebooks Recap

In Übungen 2 und 3 haben sie bereits gelernt, wie man Python-Code in Jupyter Notebooks ausführt. Die Website kaggle.com lässt sie Code schreiben und ausführen, ohne auf der lokalen Maschine etwas zu installieren. Sie haben bereits Variablen, Schleifen und Funktionen verwendet. Dieses mal werden wir zusätzliche Python-Pakete importieren, und einen externen Datensatz hochladen. Wir werden einen *Non-Small-Cell Lung Carcinoma* (NSCLC) Datensatz analysieren. Der GEO Identifier ist GSE74706. Der Datensatz enthält Samples von Krebspatienten, sowie von gesundem Gewebe als Referenz.

(a) Machen sie sich mit der Website wieder vertraut. Loggen sie sich in ihren bereits erstellten Account ein, oder erstellen sie einen neuen mit ihrer Email-Adresse oder Google-Account, und erstellen sie eine neues Notebook für diese Übung.

(b) Benennen sie das Notebook um in "Tutorial 9", und löschen sie die Zelle mit Beispielcode, die sich bereits im Notebook befindet. Erstellen sie eine neue Zelle, schreiben sie dort "2+2", udn führen sie aus. Wenn alles funktioniert wie es soll, kann man nun das Ergebnis unter der Zelle sehen.

(c) Klicken sie auf **+Add data** oben rechts. Wählen sie *Upload* oben rechts in dem Fenster, neben dem X. Wählen sie das zweite Symbol von oben auf der linken Seite. Nennen sie ihren Datensatz *lungcancer*. Kopieren die die folgende URL und fügen sie unter **Remote Files** ein:

https://sbcb.inf.ufrgs.br/data/cumida/Genes/Lung/GSE74706/Lung_GSE74706.csv

Klicken sie auf **+Add Remote Files**, dann auf **Create**. Der Upload wird ungefähr 1-2 Minuten dauern.

Alternativ können sie den Datensatz auch über den Link herunterladen, und ihn manuell in das Notebook hochladen.

(d) Nun werden wir den Datensatz in einen *DataFrame* einlesen, eine Datenstruktur die von dem *pandas* Paket für Python bereit gestellt wird. Ein DataFrame ist im Prinzip eine Tabelle mit Reihen und Spalten, ähnlich zu einer Excel-Tabelle oder einer SQL Datenbank. Hier ist eine Anleitung zu DataFrames.

(1) Zuerst müssen sie den Speicherort finden den die Datei auf dem Server hat. Oben rechts ist ein Ordner mit dem Name *Input*. Der Datensatz ist gespeichert unter *input* → *lungcancer* → *Lung_GSE74706.csv*. Hier könne sie den Pfad der Datei kopieren.

(2) Diese Zelle lädt die *.csv* Datei in einen DataFrame mit dem Name *lungcancer_df*:

```
import pandas as pd
file_path = "../input/lungcancer3/Lung_GSE74706.csv"
lungcancer_df = pd.read_csv(file_path)
```

(3) Erstellen sie eine neue Zelle, schreiben sie nur *lungcancer_df*, und führen sie sie aus. Nun wird die Tabelle darunter angezeigt. Wie viele Zeilen und Spalten hat sie?

(e) Als nächstes werden wir die Daten etwas aufräumen. In einem ersten Schritt speichern wir die *type*-Spalte (also ob es sich um ein Krebs- oder Referenz-sample handelt) in einer separaten Liste. Als nächstes erstellen wir einen DataFrame der nur die Genexpression enthält, also aus dem die Spalten *samples* und *type* entfernt sind. Erstellen sie eine Zelle die so aussieht:

```
labels_df = lungcancer_df["type"]
features_df = lungcancer_df.drop(["type", "samples"], axis=1)
```

(f) Schlussendlich werden wir einige Statistiken für die Daten berechnen, um einen Überblick zu bekommen.

(1) Zählen sie wie oft jedes label in dem Datensatz vorkommt:

```
labels_df.value_counts()
```

Wie *normal* und *NSCLC* (Lungenkrebs) samples befinden sich in dem Datensatz?

(2) Berechnen sie Statistiken für die Genexpressions-Daten:

```
features_df.T.describe()
```

Ein normalisierter Datensatz hat die Eigenschaft dass die Mittelwerte (mean) und die Standardabweichung (std) zwischen den Samples sehr ähnlich sind. Ist der Datensatz gut genug normalisiert oder ist eine Normalisierung notwendig?

(3) Erstellen sie ein Histogramm der durchschnittlichen Expression der Gene:

```
features_df.mean().hist()
```

Beschreiben sie kurz was sie auf dem Plot sehen.

Have fun!