HUMBOLDT-UNIVERSITÄT ZU BERLIN

Exposé

zur Masterarbeit

Entwicklung und kritische Bewertung eines Frameworks zur Bestimmung der Ähnlichkeit von pankreatischen neuroendokrinen Neoplasien zu Zellen in bekannten Differenzierungsstadien

Autor: Jan-Niklas Rössler
jan-niklas.roessler@hu-berlin.de

Gutachter: Prof. Dr. Ulf Leser
Prof. Dr. Rosario M. Piro

Betreuer: Raik Otto

14. November 2018

1 Motivation und Einführung

Mittels Analyse des Transkriptoms ist es durch bioinformatische Algorithmen möglich, die Ähnlichkeit von Tumoren und Karzinomen zu Zellen in bekannten Differenzierungsstadien zu bestimmen [1]. Für Früherkennung und personalisierte Therapie ist eine solche Quantifizierung klinisch relevant [2]. Da bisher kein publizierter Framework zur Messung dieser Ähnlichkeiten für pankreatische neuroendokrine Neoplasien (PanNEN) bekannt ist, soll in dieser Masterarbeit ein solcher entwickelt und kritisch bewertet werden.
Dem zu entwickelnden Framework liegt die durch Publikationen gestützte [3, 4] Hypothese zu Grunde, dass das Expressions-Profil von Tumoren Ähnlichkeiten zu dem Expressions-Profil von Zellen in bekannten Differenzierungsstadien zeigt. Mittels einer Metrik über die Transkriptome soll die Ähnlichkeit zwischen der Genexpression von PanNEN und der Genexpression von Stammzellen, pankreatischen Progenitorzellen und ausdifferenzierten pankreatischen Zelltypen quantifiziert werden.
Riester *et al.* konnten zeigen, dass die Pearson-Korrelation ein sinnvolles Maß für die Quantifizierung von Ähnlichkeiten zwischen Tumoren und Stammzellen ist. Newman *et al.* [5] haben in ihrer Studie gezeigt, dass es durch Dekonvolution von Expressionsdaten [6] mit Hilfe einer *Support Vector Regression* (SVR) möglich ist, Ähnlichkeiten in der Genexpression zwischen Zell-Konvoluten aus gesundem Gewebe zu messen.
In dieser Arbeit soll die Ähnlichkeit von PanNEN zu Differenzierungsstadien wie bei Newman *et al.* durch eine Dekonvolution mittels einer SVR bestimmt werden. Bisher wurde nicht gezeigt, dass dies bei Krebszellen möglich ist. Zudem werden im Gegensatz zu Riester *et al.* nicht nur paarweise Messungen durch Betrachtung der Korrelation einzelner Genexpressions-Profile vorgenommen, sondern es wird durch die Dekonvolution die Gesamtheit aller Samples betrachtet. Hierdurch entsteht ein Informationsgewinn.
Im Kontext der Transkriptom-Analyse versteht man unter der Dekonvolution von Genexpressionen die Bestimmung der relativen Anteile der Zelltypen, die im untersuchten Gewebe vorhanden sind. In Matrixnotation lässt sich das Dekonvolutionsproblem wie folgt darstellen:

$$T = C \times p + \epsilon$$

Wobei T die Expressionsmatrix eines Konvoluts darstellt (zum Beispiel erhoben durch RNA-Sequenzierung), C die Zelltyp-spezifischen Genexpressionen (Zellen in bekannten Differenzierungsstadien) und p ein Vektor mit den gesuchten Anteilen der Zelltypen im Konvolut.
Aufgabe des Frameworks ist es, mittels einer SVR eine Dekonvolution von Transkriptomdaten (RNA-Sequenzierung (RNA-Seq), *single-cell* RNA-Sequenzierung (scRNA), Microarray) von PanNEN zu ermöglichen und darüber eine Ähnlichkeitsmessung zwischen Differenzierungsstadien und PanNEN zu etablieren. Um Datenheterogenität, sowie technische und biologischen Artefakte auszuschließen, sollen die Ergebnisse durch bioinformatische Methoden auf ihre Signifikanz und Robustheit überprüft werden.

2 Zielstellung der Masterarbeit

Das Ziel der Masterarbeit ist es, einen bioinformatischen Framework zu entwickeln und durch ein Benchmark kritisch zu bewerten. Die Masterarbeit war erfolgreich, wenn durch Benchmarks gezeigt werden konnte, ob eine Dekonvolution der Expressionsdaten für eine Ähnlichkeitsmessung zwischen PanNEN und Zellen in bekannten Differenzierungsstadien geeignet ist oder nicht.
Durch eine Kreuzvalidierung soll die Eignung der Expressionsdaten untersucht werden, die für das Training der SVR verwendet werden. Anschließend soll mittels Bootstrapping und Perturbation die Robustheit der Dekonvolution anhand eines synthetischen Goldstandards analysiert werden. In einem weiteren Benchmark soll die Schwankung wichtiger Zielgrößen der Dekonvolution in Abhängigkeit steigender Störfaktoren bestimmt werden, um biologisch sinnvolle Signifikanz-Schwellen zu ermitteln. Nach erfolgreichen Benchmarks sollen Ähnlichkeiten für PanNEN bestimmt werden und die Ergebnisse durch einen Abgleich mit klinischen Metadaten validiert werden.
Der Framework wird in Form eines genererischen `R`-Packages entwickelt [7], mit Genexpressionsdaten von PanNEN als *proof of concept*.

3 Methodik

3.1 Transkriptomdaten und Markergene

Es werden Transkriptomdaten von Zellen in bekannten Differenzierungsstadien und von PanNEN benötigt. Die Expressionsdaten der Zelltypen dienen dabei als Referenz für die Dekonvolution, indem sie als Trainingsdaten für die SVR verwendet werden. Die Expressionsdaten von PanNEN werden zum Testen der Ähnlichkeitsmessung benötigt.

Tab. 1: Bereits identifizierte Transkriptomdaten. Es ist angegeben, welcher Studie die Daten entnommen sind, mittels welcher Technologie sie erhoben wurden, aus welchem Organismus sie stammen und ob es Daten für ein Differenzierungsstadium oder PanNEN sind. Die Daten von Grötzinger *et al.* sind noch nicht publiziert.

Autor	Technologie	Organismus	Typ
Baron *et al.* [9]	scRNA	*Homo sapiens*	Ausdifferenziert
Muraro *et al.* [10]	scRNA	*Homo sapiens*	Ausdifferenziert
Segerstolpe *et al.* [11]	scRNA	*Homo sapiens*	Ausdifferenziert
Lawlor *et al.* [8]	scRNA	*Homo sapiens*	Ausdifferenziert
Yan *et al.* [12]	scRNA	*Homo sapiens*	Stammzellen
Stanescu *et al.* [13]	scRNA	*Mus musculus*	Progenitorzellen
Spagnoli *et al.* [14]	RNA-Seq	*Mus musculus*	Progenitorzellen
Sadanandam *et al.* [15]	Microarray	*Homo sapiens*	PanNEN
Scarpa *et al.* [16]	RNA-Seq	*Homo sapiens*	PanNEN
Grötzinger *et al.*	RNA-Seq	*Homo sapiens*	PanNEN

Nach bereits erfolgter Sichtung sind erste Daten für Stammzellen, pankreatische Progenitorzellen und ausdifferenzierte pankreatische Zelltypen vorhanden (siehe Tabelle 1). Da aber, insbesondere für Stammzellen und Progenitorzellen, noch nicht ausreichend Daten vorhanden sind, sollen durch Literaturrecherche weitere Studien in das Trainingsset integriert werden. Für PanNEN sind bereits ausreichend Daten aus dem MapTorNet-Projekt vorhanden.
Außerdem müssen für die Dekonvolution Markergene definiert werden, welche durch ihre spezifische Expression eine Signatur für die Zellen in bekannten Differenzierungsstadien darstellen. Bei der Dekonvolution wird die Menge aller betrachteten Gene auf diese Untermenge von Markergenen reduziert, da diese den größten Informationsgehalt für eine Auftrennung bieten (vergleiche Newman *et al.* [5]). Ideale Markergene sind möglichst nur in einem spezifischen Zelltyp exprimiert und zeigen eine robuste Expression in biologischen Replikaten. Solche Gene sollen durch Literaturrecherche gefunden (zum Beispiel Lawlor *et al.* [8]) oder durch eine Analyse der differentiellen Genexpression ermittelt werden. Eine optimales Set von Markergenen soll als Teil des Benchmarks identifziert werden.

3.2 Generierung von synthetischen Expressionsdaten

Für das Benchmark sollen synthetische Expressionsdaten erzeugt werden, die als Goldstandard genutzt werden können, um den Framework zu benchmarken. Dafür sollen Positiv- und Negativkontrollen generiert werden.
Die Positivkontrollen werden basierend auf den Trainingsdaten erstellt, da für diese Samples bekannt ist in welchem Differenzierungsstadium sie sich befinden. Die Simulation konzentriert sich dabei auf die Markergene, die jedes Differenzierungsstadium charakterisieren. Diese spezifische Gen-Signatur muss bei den Positivkontrollen erhalten bleiben, um zu gewährleisten, dass eine auf Markergenen basierende Ähnlichkeitsmessung für diese Samples erfolgreich ist.
Für die Negativkontrollen können entweder Expressionswerte generiert werden, die einer zufälligen Verteilung folgen, sodass keinerlei Signatur zu erkennen ist. Oder es werden Samples aus einem anderen Gewebetyp verwendet, auf dem die SVR nicht trainiert wurde. Für Negativkontrollen sollten somit keine Ähnlichkeiten gemessen werden können.
Für die Simulation kann das `R`-Paket `powsimR` [17] verwendet werden. Oder es wird ein eigener Algorithmus erstellt, der diese Aufgabe übernimmt. Da die Genexpressionen von RNA-Seq-Sampels durch eine negative Binomialverteilung modelliert werden können, soll diese Verteilung auch hier für die Simulation verwendet werden.

3.3 Benchmarks

3.3.1 Kreuzvalidierung der Trainingsdaten

Durch eine Kreuzvalidierung soll die Qualität der gewählten Trainingsdaten analysiert werden. Es soll untersucht werden, ob sich Samples, die sich im gleichen Differenzierungsstadium befinden, gegenseitig zerlegen lassen. Dafür soll immer ein Teil der Samples für

das Training der SVR verwendet werden und für den verbleibenden Teil soll die korrekte Dekonvolution getestet werden. Die Trainingsdaten werden nach ihrer Studienzugehörigkeit aufgeteilt und zuerst soll jede Studie gegen sich selbst getestet werden. Anschließend dient jede Studie einmal als Trainings-Partition und die verbleibenden Studien bilden reihum die Test-Partition.

3.3.2 Beurteilung der Robustheit

In einem nächsten Schritt soll die Robustheit der Dekonvolution untersucht werden. Hierfür wird eine Kreuzvalidierung auf dem synthethischen Goldstandard durchgeführt. Für diesen Fall werden nur die Positivkontrollen verwendet.
Die Expressionsdaten werden zuerst durch Störterme perturbiert, die mittels einer negativen Binomialverteilung generiert werden und auf die Expressionswerte addiert werden. Durch Bootstrapping werden dann wiederholt Markergene aus allen Sampels entfernt und in jeder Iteration zufällige Partitionen für die Kreuzvalidierung erstellt. Die SVR wird auf den gestörten Trainingsdaten trainiert und dann mit Hilfe der Testdaten analysiert, inwiefern die Störung die Ergebnisse beeinflusst.

3.3.3 Beurteilung der Schwankung von p-Wert und Ähnlichkeitswert

Für die Dekonvolution mit Hilfe einer SVR lässt sich ein p-Wert berechnen, indem eine Nullverteilung für die berechneten Ähnlichkeiten bestimmt wird (vergleiche Newman *et al.* [5]). In einem weiteren Benchmark soll die Schwankung dieses p-Werts und des Ähnlichkeitswerts analysiert werden, um biologisch sinnvolle Signifikanz-Schwellen für diese Zielgrößen zu definieren.
Das Benchmark wird auf dem synthethischen Goldstandard mit Negativ- und Positivkontrollen ausgeführt. Zuerst werden die Samples ohne Störung des Inputs dekonvoluiert und mittels einer *receiver-operating-characteristic*-Kurve (ROC-Kurve) wird ein Cutoff für den p-Wert bestimmt, sodass sich eine optimale Zuordnung der Kontroll-Samples in signifikante und insignifikante Ähnlichkeiten ergibt.
Anschließend wird die Heterogenität eines Tumor-Sample simuliert, indem die Kontrollen durch Bootstrapping oder Perturbation gestört werden. Die perturbierten Samples werden dekonvoluiert und mittels der ROC-Kurve wird erneut ein optimaler Cutoff für den p-Wert ermittelt. Diese Analyse wird mehrfach wiederholt, wobei die Intensität der Störung bei jeder Wiederholung in einem festgelegten Intervall gesteigert wird.
Neben der Schwankung des p-Werts soll danach auf gleiche Weise die Schwankung der berechneten Ähnlichkeitswerte analysiert werden. Damit soll es durch dieses Benchmark möglich sein, zu beurteilen, ab wann den berechneten p-Werten und Ähnlichkeitswerten vertraut werden kann und ab wann eine Dekonvolution korrekt ist.

3.4 Analyse der Expressionsdaten von PanNEN

Nach dem Benchmark auf dem synthethischen Goldstandard soll die Ähnlichkeitsmessung anhand von Expressionsdaten von PanNEN durchgeführt werden. Die zuvor er-

mittelten Cutoffs für p-Wert und Ähnlichkeit sollen genutzt werden, um festzulegen, ab wann ein Sample einem Differenzierungsstadium zugeordnet werden kann. Die berechneten Ähnlichkeiten sollen validiert werden, indem die Ergebnisse mit klinischen Metadaten abgegichen werden, die unabhängig erhoben wurden.
Die berechneten Ähnlichkeiten sollen auf Artefakte überprüft werden, indem die Markergene der Differenzierungsstadien betrachtet werden. Es soll untersucht werden, ob Ähnlichkeitswerte nur auf einzelne Markergene zurückzuführen sind und ob dies biologisch sinnvoll ist. In solchen Fällen sollen diese Markergene aus der Analyse entfernt werden und anschließend soll überprüft werden, ob die berechneten Ähnlichkeiten stabil bleiben. Im Fall von Insulinomen wäre es zum Beispiel biologisch sinnvoll, dass eine hohe Ähnlichkeit zum ausdifferenzierten Stadium von einer starken Expression des Insulin-Gens herrührt [18].
Weiterhin kann es vorkommen, dass es für manche Samples nicht möglich ist signifikante Ähnlichkeiten zu berechnen. Dies kann darauf zurückzuführen sein, dass die Wahl der Trainingsdaten nicht optimal war, da besipielsweise die Bandbreite der gewählten Differenzierungsstadien nicht ausreichend war. Samples ohne Ähnlichkeiten sollen tiefergehend analysiert werden, um Markergene zu identifzieren, die in dieser Gruppe eine konsistente Expression zeigen. Dadurch kann ein neues Differenzierungsstadium definiert werden, was schließlich eine Ähnlichkeitsmessung für diese Samples möglich machen würde.
Außerdem sollen die berechneten Ähnlichkeiten mit den a priori bekannten Subklassifikationen der PanNEN-Samples (neuroendokrine Tumore (NET) und neuroendokrine Karzinome (NEC) nach Klassifikation der WHO [19]) abgeglichen werden. Es soll untersucht werden, ob die Ähnlichkeiten zu Differenzierungsstadien mit den Subtyp-Zugehörigkeiten korrelieren. Dies kann es ermöglichen, die Subklassifikation eines Sampel durch die Ähnlichkeitsmessung mittels einer SVR zu bestimmen.

Referenzen

[1] Markus Riester, Hua-Jun Wu, Ahmet Zehir, Mithat Gönen, Andre L. Moreira, Robert J. Downey, and Franziska Michor. Distance in cancer gene expression from stem cells predicts patient survival. *PLOS ONE*, 12(3):1–17, 03 2017.

[2] Igor Shats, Michael L. Gatza, Jeffrey T. Chang, Seiichi Mori, Jialiang Wang, Jeremy Rich, and Joseph R. Nevins. Using a stem cell–based signature to guide therapeutic selection in cancer. *Cancer Research*, 71(5):1772–1780, 2011.

[3] K. Eun, S. W. Ham, and H. Kim. Cancer stem cell heterogeneity: origin and new perspectives on CSC targeting. *BMB Rep*, 50(3):117–125, Mar 2017.

[4] A. L. Moreira, M. Gonen, N. Rekhtman, and R. J. Downey. Progenitor stem cell marker expression by pulmonary carcinomas. *Mod. Pathol.*, 23(6):889–895, Jun 2010.

[5] A. M. Newman, C. L. Liu, M. R. Green, A. J. Gentles, W. Feng, Y. Xu, C. D. Hoang, M. Diehn, and A. A. Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nat. Methods*, 12(5):453–457, May 2015.

[6] D. Venet, F. Pecasse, C. Maenhaut, and H. Bersini. Separation of samples into their constituents using gene expression data. *Bioinformatics*, 17(suppl_1):S279–S287, 2001.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[8] N. Lawlor, J. George, M. Bolisetty, R. Kursawe, L. Sun, V. Sivakamasundari, I. Kycia, P. Robson, and M. L. Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes. *Genome Res.*, 27(2):208–222, 02 2017.

[9] Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, Allon M. Klein, Douglas A. Melton, and Itai Yanai. A single-cell transcriptomic map of the human and mouse pancreas reveals inter- and intra-cell population structure. *Cell Systems*, 3(4):346 – 360.e4, 2016.

[10] Mauro J. Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van†Gurp, Marten A. Engelse, Francoise Carlotti, Eelco J.P. de Koning, and Alexander van†Oudenaarden. A single-cell transcriptome atlas of the human pancreas. *Cell Systems*, 3(4):385 – 394.e3, 2016.

[11] A. Segerstolpe, A. Palasantza, P. Eliasson, E. M. Andersson, A. C. Andreasson, X. Sun, S. Picelli, A. Sabirsh, M. Clausen, M. K. Bjursell, D. M. Smith, M. Kasper, C. Ammala, and R. Sandberg. Single-Cell Transcriptome Profiling of Human

Pancreatic Islets in Health and Type 2 Diabetes. *Cell Metab.*, 24(4):593–607, 10 2016.

[12] L. Yan, M. Yang, H. Guo, L. Yang, J. Wu, R. Li, P. Liu, Y. Lian, X. Zheng, J. Yan, J. Huang, M. Li, X. Wu, L. Wen, K. Lao, R. Li, J. Qiao, and F. Tang. Single-cell RNA-Seq profiling of human preimplantation embryos and embryonic stem cells. *Nat. Struct. Mol. Biol.*, 20(9):1131–1139, Sep 2013.

[13] Diana E. Stanescu, Reynold Yu, Kyoung-Jae Won, and Doris A. Stoffers. Single cell transcriptomic profiling of mouse pancreatic progenitors. *Physiological Genomics*, 49(2):105–114, 2017. PMID: 28011883.

[14] Elisa Rodríguez-Seguel, Nancy Mah, Heike Naumann, Igor M. Pongrac, Nuria Cerdá-Esteban, Jean-Fred Fontaine, Yongbo Wang, Wei Chen, Miguel A. Andrade-Navarro, and Francesca M. Spagnoli. Mutually exclusive signaling signatures define the hepatic and pancreatic progenitor cell lineage divergence. *Genes & Development*, 27(17):1932–1946, 2013.

[15] Anguraj Sadanandam, Stephan Wullschleger, Costas A. Lyssiotis, Carsten Grötzinger, Stefano Barbi, Samantha Bersani, Jan Körner, Ismael Wafy, Andrea Mafficini, Rita T. Lawlor, Michele Simbolo, John M. Asara, Hendrik Bläker, Lewis C. Cantley, Bertram Wiedenmann, Aldo Scarpa, and Douglas Hanahan. A cross-species analysis in pancreatic neuroendocrine tumors reveals molecular subtypes with distinctive clinical, metastatic, developmental, and metabolic characteristics. *Cancer Discovery*, 5(12):1296–1313, 2015.

[16] A. Scarpa, D. K. Chang, K. Nones, V. Corbo, A. M. Patch, P. Bailey, R. T. Lawlor, A. L. Johns, D. K. Miller, A. Mafficini, B. Rusev, M. Scardoni, D. Antonello, S. Barbi, K. O. Sikora, S. Cingarlini, C. Vicentini, S. McKay, M. C. J. Quinn, T. J. C. Bruxner, A. N. Christ, I. Harliwong, S. Idrisoglu, S. McLean, C. Nourse, E. Nourbakhsh, P. J. Wilson, M. J. Anderson, J. L. Fink, F. Newell, N. Waddell, O. Holmes, S. H. Kazakoff, C. Leonard, S. Wood, Q. Xu, S. Hiriyur Nagaraj, E. Amato, and I. Dalai et al. Corrigendum: Whole-genome landscape of pancreatic neuroendocrine tumours. *Nature*, 550(7677):548, 10 2017.

[17] Beate Vieth, Christoph Ziegenhain, Swati Parekh, Wolfgang Enard, and Ines Hellmann. powsim: Power analysis for bulk and single cell rna-seq experiments. *bioRxiv*, 2017.

[18] William R. Burns and Barish H. Edil. Neuroendocrine pancreatic tumors: Guidelines for management and update. *Current Treatment Options in Oncology*, 13(1):24–34, Mar 2012.

[19] F.T. Bosman, World Health Organization, and International Agency for Research on Cancer. *WHO Classification of Tumours of the Digestive System*. WHO Classification of Tumours. International Agency for Research on Cancer, 2010.