



Dissertation

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)

Customized Workflow Development and
Omics Data Integration Concepts in
Systems Medicine

Promotionsgebiet Systembiologie und Bioinformatik

Fakultät für Informatik und Elektrotechnik
Universität Rostock

Markus Wolfien

Geboren am 15. Februar 1989 in Haldensleben
Matrikelnummer: 212206712
Datum der Abgabe: 17.07.2020

Betreuer

Prof. Olaf Wolkenhauer
Institut für Informatik
Lehrstuhl für Systembiologie und Bioinformatik
Universität Rostock
Universitätsplatz 1
18051 Rostock

Zusammenfassung

Hintergrund: Die ständig wachsende Menge und Vielfalt biologischer und medizinischer Daten erhöht die Komplexität der nachfolgenden Computeranalysen erheblich. Daher müssen diese Berechnungsmethoden zu umfassenden Analyse “*Workflows*” kombiniert werden, um eine nahtlose, schnelle und transparente Berechnung oder Weiterverarbeitung zu gewährleisten. Die Systemmedizin als Ansatz zur Verwendung und Integration von mathematischen und computergestützten Modellen hat bereits ein großes Potenzial zur Untersuchung komplexer Daten von molekularen bis zu organismischen Ebenen bewiesen. Daher nutzt diese interdisziplinäre Arbeit modernste integrative Ansätze, um die heterogene Datenanalyse ebenfalls zu verbessern und die translationale Forschung von Herz-Kreislauf-Erkrankungen und Herzregeneration zu unterstützen.

Methoden: Die Workflowentwicklung, hier definiert als die sukzessive Verwendung von Softwareanwendungen, erwies sich als eines der wichtigsten Konzepte für eine transparente und wiederverwendbare Datenanalyse. Insbesondere in der stark wachsenden Landschaft von Analysewerkzeugen für die RNA Sequenzierung (RNA-Seq) wird es immer wichtiger, die jeweiligen Ergebnisse so verständlich wie möglich abzurufen, um eine qualitativ hochwertige und reproduzierbare Forschung sicherzustellen. In dieser vorliegenden Arbeit wurden zahlreiche Workflows für die allgemeine Verarbeitung der RNA-Seq bzw. Einzelzellsequenzierungsexperimente (scRNA-Seq) und die nichtkodierende RNA Identifizierung entwickelt. Alle Workflows werden öffentlich über Galaxy und FairdomHub bereitgestellt. Zum Beispiel umfasst TRAPLINE neben grundlegenden Qualitätskontroll-, Normalisierungs- und differentiellen Expressionstestverfahren auch erweiterte Funktionen für die Vorhersage von mikro RNA Zielgenen, die Identifizierung von Protein-Protein Interaktionen und Anreicherungsanalysen von Genen. Weitere Anwendungen von RNA Geschwindigkeitsberechnungen, nachgeschalteten Netzwerkansätzen wie gewichteter Genkoexpression und Signalwegsanreicherungsanalysen haben sich als entscheidend für die Validierung und Evaluation der zuvor identifizierten differentiell exprimierten Transkripte erwiesen. Neben der Entwicklung und Pflege von Workflows wurde ein Beitrag zu öffentlich verfügbarem online Schulungs-

material für die Sequenzdatenanalyse geleistet, um die wissenschaftliche Ausbildung zu verbessern. Konzepte des maschinellen Lernens und univariate Metaanalysen wurden zusätzlich implementiert, um die Rolle von Zelltherapien bei der Herzregeneration zu untersuchen.

Resultate: Um tiefere Einblicke in die transkriptomischen Veränderungen der Herzregeneration zu ermöglichen, wurde ein Analyseablauf für die RNA-Seq entwickelt, kritisch bewertet und durch einen biologischen Vergleich von stammzellabgeleiteten Kardiomyozytensubtypen validiert. Die induzierten Sinusknoten, die zur Kategorie der *in vitro* generierten Herzschrittmacherzellen nach externer *Tbx3* Verabreichung und *Myh6* basierter Antibiotikaselektion gehören, ergaben die erfolgversprechendste transkriptomische Ähnlichkeit im Vergleich zu einem menschlichen Sinusknoten. Darüber hinaus wurde die RNA Expression verwendet, um das Hubgen γ 2 AMPK unabhängig über eine gewichtete Koexpressionsanalyse zu validieren, was letztendlich bestätigt durch weitere Experimente darauf hinweist, dass es für die Herzfrequenzerzeugung ein wesentlicher Faktor ist. Im Gegensatz dazu zeigten Genanreicherungsanalysen alternativer mRNA basierter Strategien zur kardialen Reprogrammierung menschlicher Stromazellen in Kardiomyozyten nur eine unvollständige Reprogrammierung ohne schlagende Schrittmacherzellen. Weiterhin wurde die scRNA-Seq angewendet, um die erste Charakterisierung eines erwachsenen Säugetierherzens zu erhalten und realistische Zelltypverteilungen in Kombination mit RNA Geschwindigkeitskinetiken bereitzustellen. Um eine generelle Antwort auf die Wirksamkeit zellbasierter Therapien zur Behandlung von Myokardschäden bei Mäusen zu geben, zeigte die durchgeführte univariate Metaanalyse und die anschließende Metaregressionsanalyse einen positiven Effekt bei der Herzregeneration. Diese präklinischen Ergebnisse konnten mit den Ergebnissen der klinischer Studie PERFECT validiert werden, in dem das angewandte Modell des maschinellen Lernens eine Biomarkersignatur identifizieren konnte. Diese Biomarker werden zur präoperativen Auswahl von reaktiven Patienten verwendet, die nach einem Myokardinfarkt eine Behandlung mit allogenen Zellen erhalten.

Auswirkung: Im Rahmen dieser Arbeit wurden acht Trainings innerhalb des Deutschen Netzwerks für Bioinformatik Infrastruktur ([de.NBI](#)) für mehr als einhundert Doktoranden und Postdocs zur RNA-Seq Datenanalyse abgehalten. Insgesamt wurden zwei Bachelor- und drei Masterarbeiten betreut. Darüber hinaus diente diese Arbeit als Grundlage für zwei internationale Patente, zwanzig von Experten begutachteten Manuskripten und erzeugte Datenressourcen mit bereits mehr als zweitausend Downloads. Die Ergebnisse führten zu drei erfolgreich finanzierten Projektvorschlägen ([de.STAIR](#), [GB-XMAP](#), [iRhythmics](#)) mit einem Gesamtfinanzierungsvolumen von mehr als drei Millionen Euro.