DISS. ETH N° 25702

Adding Another Dimension: Computational Analysis of Temporal and Integrative Datasets to Dissect Early Embryonic Developmental Deficiencies of RNAi Mutants

a thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

DANIEL SPIES

Master of Science

University of Vienna

born on 05.07.1988

citizen of

Austria

accepted on the recommendation of

Prof. Constance Ciaudo, examiner

Prof. Christian von Mering, co-examiner

Prof. Manfred Claassen, co-examiner

Prof. Uwe Ohler, co-examiner

SUMMARY

Complex regulatory networks define the exact states of every single cell from the developing embryo to the terminally differentiated neurons in the brain. While researchers developed a plethora of methods to dissect these regulatory networks, most cases only presented snapshots of single events and failed to capture the dynamical aspects of biological processes. With the advance of next generation sequencing (NGS) technologies, genome wide approaches enabled a broader evaluation, investigating interconnected processes at the same time. Improved protocols and decreasing costs now allowed scientists to inspect processes on a genome wide level in time course experiments or across a multitude of conditions on a routinely basis, which opened a new field and challenged existing analysis methods to handle the additional dimensions captured. In this thesis we investigated two approaches to handle additional temporal or conditional dimensions of NGS data.

First, we compiled a broad range of tools designed to perform RNA-seq time course differential expression analysis and performed an extensive simulation study to characterize and rank each tool and define new best practice rules that so far have not been established. Best performing tools and defined practices were then applied to two independent biological time course experiments that captured a short signaling pathway inhibition and a long-term differentiation process, respectively. Both applications were able to further dissect underlying mechanisms and identify potential regulators and crosstalk targets for the short pathway inhibition and charted the progressive development into an aberrant cell fate. Insights gained encompass the ability and limits of a single data type to capture such complex processes and how crucial both; proper experimental design and preliminary experiments are to properly specify the resolution and range of time course experiments.

Second, we investigated the role of RNA interference (RNAi) in mouse embryonic development by integrative analysis of all pathway actor mutants on RNA and protein levels. By categorizing differential candidates into RNAi pathway centric defined classes, we were able to identify underlying mechanisms that might explain the common exit from pluripotency phenotype. While this phenotype was not obvious on the transcriptional level and did only marginally translate to the protein level, integrative analysis enabled us to pinpoint tampered transition from the naïve to the primed pluripotency state of

cells, thus limiting the differentiation potential of our RNAi mutant mouse embryonic stem cells (mESCs). Furthermore, only a small fraction of consistently up-regulated genes was detected at the RNA and protein levels, thus challenging the assumed miRNA regulatory impact suggested by computational predictive models. Although, we would like to emphasize the stem cell and mutant specific characteristics, which nevertheless impact a broad filed of research using RNAi mutants for the investigation of biological hypotheses.

In summary, we present a body of work that suggests best practices and tools to analyze time series RNA-seq data and for the first time characterized mESC RNAi mutants in an integrative manner on the transcriptional and translational level, challenging the dogma of miRNA regulatory impact in a stem cell specific manner.

ZUSAMMENFASSUNG

Der Status einer jeden Zelle ist von komplexen regulatorischen Netzwerken definiert, von Anbeginn der embryonalen Entwicklung bis hin zu terminal differenzierten Neuronen. Wissenschaftler haben daher eine Vielzahl von Methoden entwickelt, um diese regulatorischen Netzwerke zu decodieren, in den meisten Fällen jedoch erlauben diese Methoden nur eine Momentaufnahme des jeweiligen Status eines biologischen Prozesses zu studieren. Mit der fortschreitenden Entwicklung der „Next Generation Sequencing“ (NGS) Methoden wurde es möglich genomweite Studien zu konzipieren, welche eine simultan breitere Erfassung sowie Evaluierung von vernetzten Prozessen ermöglicht hat. Die fortschreitende Verbesserung von experimentellen Protokollen kombiniert mit sinkenden Kosten ermöglicht es Wissenschaftlern nun auch diese komplexen Prozesse routinemäßig über mehrere Zeitpunkte hinweg zu erfassen. Dies hat nicht nur ein neues Feld von experimentellen Methoden geschaffen, sondern erfordert nun auch die analytischen Werkzeuge die für statische Analysen konzipiert worden sind zu überdenken, um die zusätzlichen Dimensionen und somit Informationen entsprechend modellieren zu können.

In dieser Doktorarbeit werden zwei Lösungsansätze untersucht um temporale und konditionsabhängige Erweiterungen in transkriptionsweiten Datensätzen zu analysieren.

Im ersten Ansatz wurden zunächst diverse publizierten Methoden für die Analyse von longitudinal erfassten RNA-seq Experimenten zusammengestellt. Diese Methoden wurden anschließend in einer ausführlichen Simulationsstudie charakterisiert um einerseits die am geeignetsten Methoden zu identifizieren sowie generelle Standards für weitere Anwendungen zu definieren. Die zweckmäßigsten Methoden der Simulation wurden auf zwei unabhängige Datensätze angewandt, wovon der erste Datensatz eine kurze Inhibition von Signalwegen repräsentiert, und der zweite Datensatz einen Differenzierungsprozess von mehreren Tagen beschreibt. In beiden Applikationen haben die ausgewählten Methoden zugrundeliegende Mechanismen und potentielle Regulatoren der beiden Prozesse identifiziert. Spezifische übergreifende Regulatoren in der Signalweg Inhibierung sowie progressive Prozesse in der abweichenden Entwicklung zu Fibroblasten in der Differenzierungsstudie konnten entdeckt werden. Es konnten maßgebliche

Erkenntnisse in beiden Datensätze hinsichtlich der Fähigkeit und Limitierung einzelner Datentypen um komplexen Prozesse zu beschreiben gewonnen werden. Ebenso wurde festgestellt, dass vorausgehende Experimente essentiell zur Planung der Zeitpunkte und deren Abstand beitragen um eine bestmögliche Abdeckung eines biologischen Prozesses gewährleisten zu können.

Im zweiten Ansatz wurde der Einfluss von RNA Interferenz (RNAi) in Maus embryonalen Stammzellen untersucht. Dazu wurden Mutanten eines jeden RNAi Ablaufs in integrativer Art und Weise auf Transkriptions- und Translationsebene untersucht. Identifizierte differentielle Kandidaten wurden in einer RNAi zentrischen Art und Weise kategorisiert. Dies ermöglichte die Identifizierung potentieller zugrundeliegenden Mechanismen des geteilten Phänotypen: die fehlgeschlagene Terminierung der Stammzell Pluripotenz. Weiterhin konnten wir feststellen, dass bisher vorgesagter regulatorischer Einfluss von miRNAs auf unsere RNAi Mutanten auf RNA Ebene nur minimal ist und kaum auf die Tranlastionsebene übertragen wird. Dennoch konnten wir eine Dysregulierung des Übergangs von einem naiven zu einem Vorläuferstadium der Differenzierung feststellen, und somit einen potentiellen Grund für das fehlende Differenzierungspotential darstellt.

Zusammenfassend stellen wir in dieser Arbeit neue Standards für die Analyse von zeitlichen RNA-seq Abfolgen vor und haben innovativ alle möglichen RNAi Mutanten integrativ auf transkriptioneller und translativer Ebene verglichen, was zu einer Infragestellung bisherigen Dogmen der miRNA Regulation in Maus Stammzellen führt.