Verständnis von zellulärern Zuständen und Zellzustandsübergängen durch integrative Analyse epigenetischer Veränderungen[1]

Dr. Michael J. Ziller [2]

Abstract: Wie kommt es, dass ein und dasselbe Genom eine derartige Vielzahl unterschiedlicher Zelltypen erzeugen kann? Die Antwort auf diese Frage liegt in der epigenetischen Regulation des Genoms. Im Rahmen dieser Dissertation wurden statistische Verfahren und neue Analyseansätze zur Auswertung von genomweiten Messungen des epigenetischen Zustands in verschiedenen Zellpopulationen entwickelt. Diese Verfahren gestatten es, die der Zellidentitätsbildung zugrunde liegenden molekularen Mechanismen besser zu verstehen und Teile der zugehörigen regulatorischen Netzwerke zu dekodieren. Auf diese Weise konnten diverse neue biologische Einsichten gewonnen und validiert werden. Darüber hinaus sind die vorgestellten Methoden in weiten Teilen der Epigenomik anwendbar und zeigen neue Konzepte zur Analyse von hochdimensionalen genomischen Daten auf [Zi14].

1 Einführung

Wie kann ein einziges Genom, das in nahezu allen Zellen eines Organismus identisch ist, eine derartige Vielfalt an hochgradig spezialisierten Zelltypen hervorbringen, wie wir sie in komplexen Organismen finden können? Dies ist eine der zentralen biologischen Forschungsfragen der letzten 100 Jahre. Dennoch sind nach wie vor viele Aspekte dieser Frage ungeklärt. In den vergangenen zehn Jahren gewonnene Erkenntnisse weisen jedoch der räumlich-zeitlichen Kontrolle der Genaktivitäten durch epigenetische Mechanismen eine zentrale Rolle zu. Letzere sind gekennzeichnet durch biochemische Prozesse die zur Ausbildung höherer Ordnungsstrukturen der DNA - wie z.B. Faserbildung mit hoher Packungsdichte - führen, ohne jedoch die DNA-Sequenz selbst zu beeinflussen. Mittlerweile ist eine Vielzahl verschiedener epigenetischer Modifikationen bekannt, die unterschiedlichen genomweiten Verteilungsmustern folgen. Jede dieser Modifikationen kann als eigene Dimension im Raum der epigenetischen Zustandsvektoren des Genoms betrachtet werden.

Epigenetische Regulationsmechanismen gestatten eine zelltyp- und zellzustandsspezifische Aktivierung oder Repression einzelner Gene und bilden somit die Grundlage für die Expression einer einzigartigen Kombination von Genen in jedem Zellzustand bzw. Zelltyp (siehe Abbildung 1).

Da in jedem einzelnen der etwa 200 Zelltypen des menschlichen Körpers nur etwa 15,000 der 40,000 Gene aktiv sind, birgt diese Art von Regulation enormes kombinatorisches

[1] Englischer Titel der Dissertation: “Dissecting cellular states and cell state transitions through integrative analysis of epigenetic dynamics”

[2] Department of Stem Cell and Regenerative Biology, Harvard University, michael_ziller@harvard.edu

Potenzial. Die Natur nutzt dieses Potenzial um eine Vielzahl unterschiedlicher zellulärer Phänotypen mit jeweils unterschiedlichen epigenetischen und Genexpressionszuständen zu erqeugen.
Eine zentrale Rolle zur Bestimmung der Genexpressions- und epigenetischen Profile einzelner Zelltypen kommt dabei Hochdurchsatz-Messverfahren wie hochgradig parallele Sequenzierung zu. Nur durch die Nutzung derartiger Verfahren ist es heute möglich, den epigenetischen Zustand des gesamten Genoms in zahlreichen epigenetischen Dimensionen sowie die Expression aller Gene in nahezu jedem Zelltyp und -zustand zu erfassen. Die

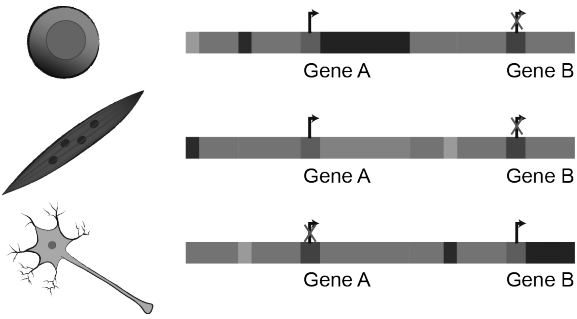

Abb. 1: Illustration der unterschiedlichen epigenetischen Zustände (farbige Abschnitte) des menschlichen Genoms (Balken) in verschiedenen Zelltypen. Je nach epigenetischem Zustand ist ein Gen aktiv (links) oder abgeschaltet (rechts).

detaillierte Charakterisierung der molekularen Profile verschiedener Zellzustände gestattet zwar deren genaue Definition, sagt aber zunächst wenig darüber aus, wie der spezifische molekulare Zustand den beobachteten Phänotyp erzeugt, etwa die Gehirn-, Blut- oder Leberzelle. Insbesondere ist unklar, warum an bestimmten Positionen im Genom im Laufe der Ausbildung unterschiedlicher Zelltypen epigenetische Veränderungen stattfinden, da viele dieser Veränderungen nicht direkt mit Genen ko-lokalisieren.
Der funktionale Anteil dieser Veränderungen wird gezielt zu genregulatorischen Elementen geleitet, deren Aktivitätszustand wiederum das Aktivitätslevel der kontrollierten Gene reguliert.Diese genregulatorischen Elemente weisen vor allem im aktiven Zustand ein charakteristisches epigenetisches Modifikationsmuster auf. Basierend auf dem epigenetischen Zustandsvektor eines jeden Elements lässt sich nicht nur der Aktivitätszustand definieren, sondern auch verschiedene Klassen von regulatorischen Elementen wie z.B. Promotoren, Enhancer und Insulator-Elemente. Zunächst ist jedoch unbekannt, wo genau diese regulatorischen Elemente im Genom lokalisiert sind. Vor diesem Hintergrund ergeben sich drei Schlüsselfragen, um die Entstehung phänotypischer Zellidentiäten zu verstehen:

1. Wie können genregulatorische Elemente, die zur Zellidentiätsbildung beitragen, identifiziert werden?
2. Auf welche Art und Weise tragen die verschiedenen jeweiligen epigenetischen Zustände des Genoms zur Entstehung einer spezifischen zellulären Identität bei?
3. Welche Mechanismen kontrollieren den epigenetischen Zustand der genregulatorischen Elemente in den verschiedenen zellulären Kontexten?

Um diese Fragen zu beantworten, stehen mittlerweile enorme Datenmengen zur Verfügung, welche den epigenetischen Zustand des gesamten Genoms in zahlreichen verschiedenen epigenetischen Dimensionen sowie über Dutzende von Zelltypen kartieren. Auch die effiziente Produktion epigenetischer Profile ist heute kein limitierender Faktor mehr und wird in großem Stil fortgesetzt.
Heute liegt die Herausforderung vielmehr in der integrierten Analyse dieser heterogenen und hochgradig komplexen Datensätze und insbesondere in der Interpretation der Resultate. Als äußerst schwierig gestaltet es sich vor allem, die Auswertungsstrategie - semantisch wie algorithmisch - so zu formulieren, dass die Analyseresultate zur Beantwortung der zentralen Fragen 2 und 3 zu biologisch relevanten und sinnvollen Ergebnissen führen. Dies ist aufgrund aufgrund der Komplexität und schieren Größe des zulässigen Suchraumes, der zahlreiche mehr oder weniger biologisch relevante Erkenntnisse enthält, eine große Herausforderung. Aus diesem Grund ist es essenziell, nicht nur effiziente Algorithmen zur Datenanalyse zu entwickeln, sondern auch den Antwortraum auf die biologischen Fragen durch das Design der analytischen Instrumente so zu strukturieren, dass am Ende gut interpretierbare und für die biologischen Fragestellungen unmittelbar relevante Ergebnisse erzeugt werden.
Aus informatischer Sicht habe ich mich genau mit diesen zwei Problemen in meiner Dissertation auseinandergesetzt: Im ersten Teil habe ich mich vorwiegend mit der Entwicklung adäquater statistischer Methoden zur grundlegenden Datenalyse befasst und im zweiten Teil auf das an der biologischen Fragestellung orientierte Design eines analytischen Frameworks zur Datenevaluation konzentriert.

2 DNA-Methylierungsveränderungen

Für verschiedene epigenetische Modifikationen ist bereits bekannt, welche Arten von genregulatorischen Elementen durch sie kontrolliert werden und auch welche Teile des Genoms mit Ihnen angereichert sind. Für die DNA-Methylierung, eine der ältesten bekannten epigenetischen Modifikationen, ist dies jedoch nur begrenzt der Fall. Insbesondere ist nicht bekannt, welche Teile des Genoms ihren DNA-Methylierungszustand als eine Funktion der Zellidentität ändern und welche konkrete Bedeutung derartige Veränderungen haben könnten.
Um diese Frage zu beantworten, haben wir das bisher größte Kompendium von genomweiten DNA-Methylierungsdatensätzen generiert. Dieses umfasst über 30 verschiedene Zelltypen und Zellzustände. Jeder dieser Datensätze enthält etwa 27 Millionen Datenpunkte, wobei jeder Datenpunkt den Prozentsatz der methylierungssensitiven Dinucleotide (CpGs) an einer Position im menschlichen Genom darstellt.
Die erste grundlegende Herausforderung bestand darin, all diejenigen CpG Dinukleotide zu ermitteln, die ihren Methylierungszustand zwischen einem beliebigen Zelltyppaar signifikant ändern. Um diesen Aspekt besser zu verstehen, ist es nützlich, den Prozess, der zur Generation der Methylierungsprozentsätze führt, zu verstehen und dann zu modellieren: Das Methylierungsmessverfahren operiert auf einer großen Population von Zellen und wählt zufällig einzelne Zellgenome aus, in denen der Methylierungszustand eines einzelnen Dinukleotids an genomischer Position x gemessen wird. (siehe Abbildung 2). Dieser

Methylierungszustand ist eine binäre Größe (methyliert oder unmethyliert). Der gleiche Prozess findet für alle CpG-Positionen im Genom parallel statt, sodass für jede Position, je nach technischer Qualität des Messprozesses, etwa 30-50 verschiedene Zellen ausgewählt werden. Oft werden unterschiedliche Positionen in den gleichen Zellen gemessen, jedoch ist, außer für nah beieinanderliegende CpGs, nicht klar, ob ihr Methylierungszustand in der gleichen Zelle gemessen wurde oder nicht.

Das Ergebnis dieses Prozesses ist für jede genomische CpG Koordinate ein binärer Vektor

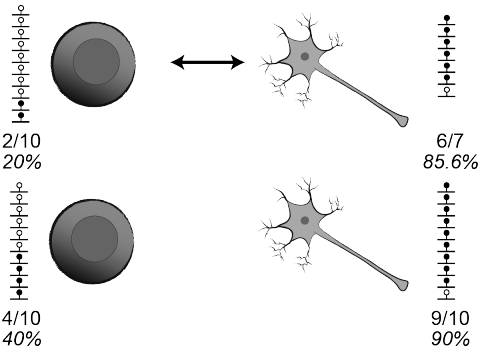

Abb. 2: Illustration der Messergebnisse des Methylierungszustands in verschiedenen Zelltypen und des resultierenden Problems zur Bestimmung differentiell methylierter genomischer Positionen. Jeder kleine *Lollipop* repräsentiert ein gemessenes CpG Dinukleotid, wobei ein ausgefüllter Kreis für ein methyliertes und ein weisser Kreis für ein unmethyliertes CpG steht. Die angegebenen Zahlen geben den Schaätzwert des kumulativen Methylierungszustands dieses CpG Dinukleotids in der gesamten Zellpopulation an.

unterschiedlicher Länge, der den Methylierungszustand aller gemessenen CpGs enthält. Die Länge des Vektors, also die Anzahl der gemessenen CpGs, wird als Coverage bezeichnet, und der Anteil der methylierten CpGs an einer Position ergibt den Methylierungsprozentsatz oder das Methylierungslevel. Dieser Prozentsatz spiegelt den Schätzwert der Anteile aller Zellen in der Population wider, die an Position x ein methyliertes CpG haben. Die Konfidenz in die Schätzung des korrekten Anteils hängt dabei von der Coverage an jeder Position ab, die in jedem Datensatz unterschiedlich ausfallen kann.

Weiterhin kann das Methylierungslevel an einer Position zwischen Zellpopulationen des gleichen Typs unter identischen Bedingungen aufgrund der inhärenten Stochastizität des biologischen Systems teilweise stark variieren. Dementsprechend muss bei der Bewertung der statistischen Signifikanz von Unterschieden im Methylierungslevel diesen beiden Unsicherheitsquellen Rechnung getragen werden.

Bisherige Verfahren im Feld haben entweder den diskreten Charakter der zugrunde liegenden Daten ignoriert und sich auf den Vergleich der Prozentsätze konzentriert oder sich ausschließlich auf die Sampling induzierte Varianz fokussiert. Wünschenswert ist jedoch ein Verfahren, das sowohl die durch den diskreten Sampling-Prozess generierte Varianz als auch die natürliche biologische Varianz zwischen Replikaten inkorporiert.

Aus diesem Grund habe ich zur Modellierung des gesamten Messprozesses ein hierarchisches Modell verwendet, das beide Aspekte berücksichtigt. Der Messprozess in einer einzelnen Population wird dabei mit einem Binomial-Modell mit unbekannter Methylierungs-

wahrscheinlichkeit beschrieben. Letztere wird dann ebenfalls als Zufallsvariable mit einer Beta-Verteilung aufgefasst. Jede gemessene Zellpopulation, also z.B. jedes biologische Replikat, besitzt ihre eigene Methylierungswahrscheinlichkeit, die aus der übergreifenden Beta-Verteilung gezogen wurde. Auf diese Weise kann sowohl der Sampling induzierten wie der biologisch erzeugten Unsicherheit Rechnung getragen werden. Beide Prozesse werden dann in einem Modell kombiniert; wobei die Beta-Verteilung als Prior des Methylierungsparameters der Binomialverteilung behandelt wird. Mittels eines empirical Bayes Verfahrens können dann die Parameter der Beta-Verteilung aus den Daten für jedes genomische CpG in einem Zelltyp bestimmt werden. Aufbauend auf diesen Ergebnissen kann dann das Methylierungslevel jedes genomischen CpGs in jedem Zelltyp mittels einer Beta-Verteilung beschrieben werden.
Um nun signifikante Unterschiede einzelner CpGs zwischen zwei Zelltypen zu ermitteln, verwende ich diese statistische Beschreibung und subtrahiere die Verteilungen der Methylierungslevel voneinander, was der Subtraktion zweier Beta-Verteilungen gleichkommt. Die resultierende Beta-Differenz-Verteilung ist in der Literatur bekannt, und es können Konfidenzintervalle und p-Werte berechnet werden. Auf diese Weise konnte ich ein statistisches Modell etablieren, das - unter Berücksichtigung von Sampling und biologischer Varianz - die Signifikanz von Methylierungsveränderungen zwischen verschiedenen Zelltypen - ermittelt.
Basierend auf diesem Modell konnte ich im nächsten Schritt einen Großteil der dynamisch methylierten CpGs im menschlichen Genom ermitteln. Vielfach treten dynamische CpGs jedoch nicht isoliert auf, sondern in räumlicher Nachbarschaft, sodass sich derartige dicht beieinanderliegende CpGs zu Gruppen oder Dynamisch Methylierten Regionen (DMRs) zusammenfassen lassen. Eine Extrapolation dieser Ergebnisse zu größeren Datensätzen legt nahe, dass etwa 25% aller CpGs im menschlichen Genom ihren Methylierungszustand im Laufe der Organisßentwicklung verändern.
Im zweiten Teil der Analyse haben wir uns der Frage nach der Funktionalität dieser Veränderungen im Allgemeinen sowie für die spezifischen untersuchten Zelltypen zugewandt. Zur genauen Charakterisierung haben wir dazu zunächst ein Kompendium von verschiedenen, bereits bekannten Annotationen über viele verschiedene Zelltypen erzeugt, das auch zahlreiche andere Arten von epigenetischen Zustandsdaten enthält.
Mittels dieses Kompendiums konnten wir zeigen, dass DMRs ein allgemeiner Marker von genregulatorischen Elementen sind. Insbesondere fallen Regionen, die zelltypspezifisch ihren Methylierungszustand verändern, mit für diesen Zelltyp kritischen genregulatorischen Elementen zusammen. Diese Beobachtung wird zusätzlich von der Erkenntnis unterstützt, dass zelltypspezifische DMRs signifikant für diejenigen genetischen Veränderungen angereichert sind, die bereits mit Krankheiten assoziiert wurden.

Zusammenfassung Der erste Teil meiner Dissertation hat sich vor allem mit den Fragen nach der Identifikation genregulatorischer Elemente befasst, die zur Zellidentitätsbildung sowie deren Funktionen für die unterschiedlichen zellulären Kontexte von Bedeutung sein können. Dazu habe ich mich vor allem auf die DNA-Methylierung konzentriert. Durch die statistische Modellierung des Messprozesses konnte ich sodann einen Großteil der dynamisch methylierten Regionen im menschlichen Genom identifizieren. Daraufhin war es

möglich, die resultierenden DMRs im zweiten Teil unserer Analyse einen Teil der biologisch relevanten Semantik der DNA-Methylierungsveränderungen zu rekonstruiert.

3 Interpretation von Zellzustandsübergängen

Der zweite Teil meiner Dissertation geht den Fragen nach, (1.) welche Mechanismen den epigenetischen Zustand der genregulatorischen Elemente in verschiedenen zellulären Kontexten kontrollieren und (2.) wie Veränderungen der epigenetischen Muster zum besseren Verständnis der genregulatorischen Logik genutzt werden können. Diese Fragen traten im Zusammenhang mit einer Studie zur in vitro-Differenzierung humaner embryonaler Stammzellen in mehrere Stufen neuraler Vorläuferzellen auf. Im Zuge einer 200-tägigen Zeitreihe wurden sechs verschiedene Zeitpunkte für eine detaillierte epigenetische Charakterisierung ausgewählt.
Da jeder Zeitpunkt durch einen einzigartigen Phänotyp wie z.B. Morphologie gekennzeichnet ist, war eine der zentralen Fragen, welche regulatorischen Mechanismen diese Phänotypveränderungen im Laufe der Zeit verursachen. Weiterhin ist es von großem Interesse zu verstehen, wie die epigenetischen Veränderungen während der Differenzierung zur Entstehung der spezifischen Zelltypen beitragen.
Aus informatischer Sicht hat man es also mit einem hochdimensionalen (gesamtes Genom) Zeitreihenproblem zu tun, für das verschiedene Arten von epigenetischen und Genexpressions-Datentypen erhoben wurden. Standardansätze im Feld befassen sich vielfach mit jedem Datentyp einzeln, um Unterschiede zwischen den verschiedenen Zeitpunkten zu ermitteln, woraufhin die epigenetisch dynamischen Regionen genauer charakterisiert werden, z.B. mithilfe diverser genomischer Annotationsbibliotheken. Andere populäre Ansätze analysieren alle Daten gemeinsam, um zunächst die Korrelationsstruktur zwischen verschiedenen Datentypen in Form eines Hidden Markov Modells zu lernen. Dabei spiegelt die Anzahl der Zustände dieses Modells die in den Daten realisierte Kombinatorik verschiedener epigenetischer Modifikationen an derselben genomischen Position wider.
Beide Ansätze haben verschiedene Nachteile hinsichtlich der hier relevanten Fragestellung: Die Kenntnis der epigenetisch dynamischen Regionen, sei es für jeden Datentyp einzeln oder integriert, ist zwar informativ, sagt aber zunächst wenig über die damit assoziierte Biologie aus. Daher ist es üblich im nächsten Schritt, alle Regionen mit einem teilweise umfangreichen Feature Vektor zu annotieren, der diverse biologisch relevante Attribute der einzelnen Regionen zusammenfasst. Basierend auf dieser Annotation kann dann nach überrepräsentierten Features zu einzelnen Zeitpunkten gefragt oder die Regionen in weitere Unterklassen unterteilt werden, die möglicherweise eine biologisch relevante Gruppierung ergeben. Derartige Standardansätze zur funktionalen Charakterisierung haben jedoch ebenfalls diverse Nachteile:

1. Sie ignorieren die spezifischen Eigenschaften der zugrunde liegenden Daten und des Kontextes.
2. Sie ignorieren die dynamische Natur vieler Regionen-Features: Obwohl eine Region zahlreiche Features aufweist, ist in den meisten zellulären Kontexten nur ein kleiner

Teil der Features relevant. Die Menge der biologisch bedeutsamen Features ist dabei allerdings eine Funktion des Zelltyps.

3. Die Resultate sind oft schwer interpretierbar, da sich oft zahllose (statistisch signifikante) Muster ohne biologisch relevante Bedeutung finden lassen.

4. Die Ergebnisse sagen zunächst wenig über übergeordnete Kontrollmechanismen aus, welche die epigenetischen Veränderungen verursacht haben könnten.

5. Die erreichbare Auflösung der Ergebnisse ist oft zu niedrig, um Aussagen über die Bedeutung einzelner identifizierter Gruppen genomischer Regionen für den jeweiligen Zelltyp zu machen.

Um diesen Problemen im Kontext der epigenetischen Veränderungen zwischen verschiedenen Zelltypen zu begegnen, habe ich in meiner Dissertation einen Modell-basierten Ansatz zur Analyse und Interpretation derartiger Daten entwickelt. Das verwendete Modell der epigenetischen Regulation genomischer Regionen orientiert sich dabei direkt an den aus der Biologie bekannten zellulären Kontrollstrukturen und nutzt somit biologisches Wissen, um den Suchraum zu strukturieren. Damit lässt sich die zu lösende Fragestellung auf ein Inferenzproblem zurückführen. Auf diese Weise werden die Ergebnisse der Datenanalyse biologisch direkt interpretierbar.
Das verwendete Modell macht die biologisch motivierte und weitgehend gültige Annahme, dass epigenetische Veränderungen an genomischen Regionen von der differenziellen Aktivität übergeordneter Entitäten, wie etwa Transkriptionsfaktoren, verursacht werden. Weiterhin machen wir die Annahme, dass derartige Faktoren, die zur Generierung eines Zelltyps funktional relevant sind, auch einen großen Teil der epigenetischen Unterschiede zwischen zwei Zeitpunkten verursachen. Welche Faktoren dies sind, ist jedoch unbekannt.
Basierend auf diesen Annahmen konstruieren wir ein lineares Modell, in dem der zellzustandsspezifische epigenetische Zustand als quantitative Größe eine Funktion der zelltzustandspezifischen Aktivität von Transkriptionsfaktoren ist. Der epigenetische Zustand aller genomischer Regionen ist für verschiedene epigenetische Dimensionen bekannt und eine kontinuierliche Größe. Zunächst behandeln wir jede dieser Dimensionen seperat.
Um genomische Regionen mit Transkriptionsfaktoren zu assoziieren, nutzen wir aus, dass die meisten Faktoren bekannte spezifische Basensequenzen im Genom präferiert binden. Dementsprechend verwenden wir einen Katalog bekannter Bindungssequenzen, sogenannter Motive, für über 520 Faktoren und ermitteln für jede epigenetisch dynamische Region die Faktoren, die potenziell dort binden könnten, basierend auf einer Sequenz-Mustersuche. Diese Analyse resultiert aus einer quantitativen Konnektivitätsmatrix zwischen genomischen Regionen und Transkriptionsfaktoren. Jede Zelle dieser Matrix enthält die Wahrscheinlichkeit, dass ein spezifischer Transkriptionsfaktor diese Region binden könnte.
Diese Wahrscheinlichkeit fußt jedoch ausschließlich auf der übereinstimmung des korrespondierenden Motivs mit Sequenzen, die in der genomischen Region vorhanden sind. Das bedeutet, je größer die genomische Region ist, desto mehr unterschiedliche Motive können gefunden werden. Dies wiederum führt zu einer dicht besetzen Konnektivitätsmatrix, was die Analyse schwierig macht und die Interpretierbarkeit stark herabsetzt, da aus der Biologie bekannt ist, dass nur ein kleiner Teil der auftretenden Motive auch funktional ist. Weiterhin ist bekannt, dass für eine große Klasse epigenetischer Modifi-

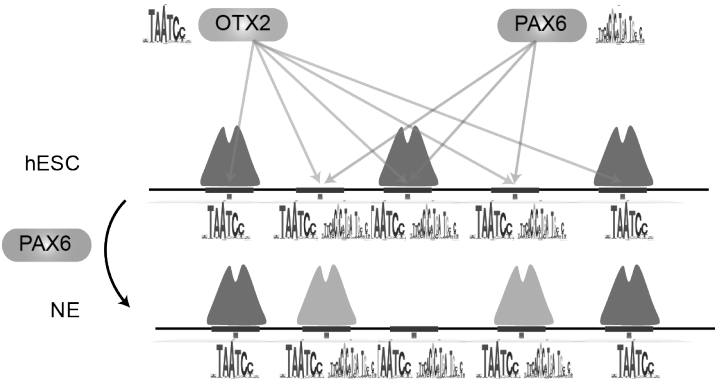

Abb. 3: Schematische Darstellung des TERA-Konzepts: Genregulatorische Elemente (lila) entlang des Genoms haben unterschiedliche epigenetische Aktivitätslevel (grüne Dreiecke) in verschiedenen Zelltypen (humane embryonale Stammzellen (hESC, oben) oder Neuroepithelzellen (NE, unten)). Die einzelnen Elemente weisen Bindugnsstellen für verschiedene Transkriptionsfaktoren auf (graue Kästchen und Motive, z.B. OTX2 und PAX6). Der TERA-Algorithmus bestimmt nun die Transkriptionsfaktoren, deren differentielle, vorhergesagte Bindungsaktivität am besten die Unterschiede in den Genomweiten epigenetischen Aktivitätsprofilen zwischen zwei Zelltypen erklärt.

kationen lokale minima in den Modifikationsdichteprofilen über genomische Regionen mit ansonsten hoher epigenetischer Modifikationsdichte bevorzugt mit Transkriptionsfaktorbindung zusammenfallen. Um dementsprechend die Falsch-Positivrate zu reduzieren, nutzt der Motiv-Identifikationsalgorithmus diese biologische Erkenntnis aus und konzentriert sich nur auf diesen kleinen Bereich innerhalb von Regionen, die epigenetische Modifikationen aufweisen. Die daraus resultierende Konnektivitätsmatrix ist sehr dünn besetzt, da im Schnitt nur fünf bis zehn Faktoren eine einzelne Region potenziell binden.

Damit sind alle wesentlichen Komponenten des Modells ermittelt. Wir beschreiben den epigenetischen Zustand der epigenetisch dynamischen Regionen als lineare Funktion der an der jeweiligen Region potenziell bindenden Transkriptionsfaktoren. Deren Aktivität ist jedoch unbekannt. Um diese zu bestimmen, betrachten wir die zellzustandsspezifische Aktivität der Transkriptionsfaktoren als Unbekannte und machen gleichzeitig die biologisch motivierte Annahme, dass die verschiedenen Faktoren zu regulatorischen Modulen zusammengefasst werden können, wobei jedes Modul mehr als einen Faktor enthalten kann. Mit dieser Formulierung und den zusätzlichen Annahmen kann das Problem effizient mit dem etablierten Partial Least Square (PLS) Verfahren und dem SIMPLS Algorithmus gelöst werden.

Dieses Verfahren liefert eine Organisation der Transkriptionsfaktoren in regulatorische Module und gestattet es gleichzeitig, genomische Regionen einzelnen Modulen zuzuordnen. Weiterhin kann für jeden Transkriptionsfaktor ermittelt werden, wie gut dieser das epigenetische Profil eines Zellzustands erklären kann. Die Differenz dieser epigenetisch inferierten Transkriptionsfaktoraktivitäten zwischen zwei Zellzuständen kann man dann als die Kapazität eines Faktors auffassen, die Unterschiede in den epigenetischen Profilen dieser Zellzustände zu erzeugen.

Ordnet man die Faktoren nach dieser epigenetischen Remodellierungsaktivität bei einem Vergleich zweier Zellzustände, so lässt sich die resultierende Rangordnung als Gewichtung der Faktoren hinsichtlich ihrer Relevanz für den Zellzustandsübergang interpretieren. Auf diese Art und Weise lassen sich nun epigenetische Veränderungen im Laufe der Differenzierungszeitreihe aus der Perspektive der differenziellen Aktivität von Transkriptionsfaktoren betrachten.
Weiterhin gestattet es eine derartige Ordnung der Faktoren, Schlüsselfaktoren der Zellzustandsübergänge zu identifizieren. Dieses Verfahren lässt sich für jede epigenetische Dimension einzeln anwenden und auch die Resultate sind einzeln interpretierbar. Manchen Faktoren sind mehr in die Veränderungen einzelner Modifikationen involviert als andere. Um die biologische Plausibilität von identifizierten Schlüsselfaktoren zu erhöhen, haben wir im nächsten Schritt die Inferenzresultate basierend auf den einzelnen epigenetischen Dimensionen zusammengefasst und eine auf diesen Ergebnissen beruhende integrierte Rangordnung erstellt. Dies führt dazu, dass Faktoren, die hohe Relevanz in der Remodellierung mehrerer epigenetischer Modifikationen haben, einen höheren Rang erhalten als solche, die nur für eine Modifikation von Bedeutung sind. Auf diese Weise ist es möglich, die heterogenen Datensätze über viele verschiedene epigenetische Dimensionen leicht interpretierbar miteinander zu kombinieren.
Im letzten Schritt haben wir uns schließlich der überprüfung unserer Resultate und Modelle zugewandt. Dabei war es ausdrücklich nicht das Ziel, unser Modell oder einzelne Annahmen direkt zu validieren, da diese eindeutig die komplexen Zusammenhänge stark vereinfachen und sicherlich nur in einem Teil der Fälle zutreffen. Vielmehr war es unsere Zielsetzung, die Nützlichkeit unserer Methode zur Generierung experimentell direkt überprüfbarer Hypothesen zu demonstrieren.
Zu diesem Zweck haben wir die resultierende Rangordnung sowohl mit bereits bekannten Schlüsselfaktoren der einzelnen Differenzierungsstufen abgeglichen als auch durch umfassende Perturbationsexperimente an jeder einzelnen Differenzierungsstufe überprüft. Beide Verifikationsverfahren zeigen exzellente Validierungsraten und legen damit nahe, dass unsere Methode zur Identifikation von Schlüsselregulatoren zellulärer Zustände und Zellzustandsübergänge geeignet ist.
Diese Methode ist jedoch keineswegs auf unsere Zeitreihendaten beschränkt, sondern kann zur Identifikation von Transkriptionsfaktoren verwendet werden, deren Aktivität die epigenetischen Unterschiede zwischen zwei beliebigen Zellzuständen erklären können. Neben einem Beitrag zum Verständnis der den verschiedenen Zellzuständen zugrunde liegenden molekularen Mechanismen hat unser Verfahren auch potenziell Relevanz für die gezielte Erzeugung von spezifischen zellulären Zuständen. So wäre es möglich, durch die künstliche überexpression eines Faktorcocktails Zelltyp A in Zelltyp B zu konvertieren. Eine Anwendung in diesem Bereich erscheint vielversprechend.

4 Zusammenfassung

Im Rahmen dieser Dissertation wurden drei grundlegende offene Fragen zur Ausbildung zellulärer Identität behandelt und dabei gleichzeitig neue informatische Konzepte und statistische Methoden zu deren Untersuchung entwickelt. Die Entwicklung dieser Konzepte

orientiert sich direkt an einem vereinfachten Modell der biologischen Prozesse. Ziel war es, die biologischen Fragen informatisch so abzubilden, dass die Analyseresultate direkt biologisch interpretierbar und gleichzeitig biologisch sinnvoll sind. Im Kontext von hochdimensionalen und heterogenen Datensätzen aus dem Genomics-Feld erweist sich diese Strategie als vielversprechend, da der Suchraum von interessanten Mustern auf diese Weise stark vorstrukturiert wird, um biologisch relevante Ergebnisse zu erzeugen. Diese Art der Problembehandlung setzt den Schwerpunkt auf die Entwicklung von Modellbildungsstrategien und versucht, wenn immer möglich, bereits existierende Algorithmen zur Lösung der einzelnen Teilprobleme zu verwenden. Nur dann, wenn keine adäquaten Algorithmen oder statistischen Modelle existieren, wurden ebenfalls modellorientiert angepasste Verfahren entwickelt. Diese Art der hochgradig integrierten Entwicklung von Analyse- und Auswertungsinstrumenten zeigt einen vielversprechenden gemeinsamen Weg der Lebenswissenschaften und Informatik auf.

Literaturverzeichnis

[Zi14] Ziller, Michael J.: Dissecting cellular states and cell state transitions through integrative analysis of epigenetic dynamics. Dissertation, Universität Tübingen, 2014.

Michael J. Ziller wurde am 17.07.1983 in Hamm (Westfalen) geboren und studierte von 2003-2010 Bioinformatik und Physik an der Eberhard-Karls-Universität Tübingen. Nach dem Erwerb seines Bioinformatik- (2009) und Physik-Diploms (2010) wechselte er für seine Dissertation als Visiting PhD Student an die Harvard Universität in Cambridge, USA. Seine Dissertation wurde dabei von Prof. Dr. Oliver Kohlbacher (Universität Tübingen) und Prof. Dr. Alexander Meissner (Harvard Universität) betreut. Seit dem Abschluss seiner Promotion im September 2014 an der Universitä Tübingen setzt er als Postdoc seine Projekte im Labor von Prof. Dr. Alexander Meissner fort.