Methods for DNA Methylation Sequencing Analysis and their Application on Cancer Data

Helene Kretzmer

Zusammenfassung

Das grundlegende Thema dieser Arbeit ist die Entwicklung von Hilfsmitteln für die Analyse von DNA-Methylierungsdaten, sowie deren Anwendung auf eine große Anzahl von Bisulfit sequenzierter Proben. DNA-Methylierung ist eine der wichtigsten epigenetische Modifikationen. Ungewöhnlichen Veränderungen sind mit einer Vielzahl von Krankheiten assoziiert, insbesondere mit kanzerogener Entwicklung von Geweben. Um die DNA-Methylierung zu sequenzieren, sind spezielle Techniken erforderlich.

Zu Beginn wird das Bisulfit-Analyse-Toolkit `BAT` vorgestellt, das entwickelt wurde, um eine schnelle Analyse der Bisulfit behandelten Sequenzierungsdaten zu ermöglichen. Es deckt alle Schritte der Verarbeitung von rohen Sequenzdaten bis hin zur Detektion differentieller DNA-Methylierung ab. Zudem wird die DNA-Methylierung mit Genexpressionsdaten integriert indem korrelierende Regionen berechnet werden.

Zweitens wird ein neuer Algorithmus, `metilene`, für die Berechnung von differentiell methylierter Regionen (DMRs) zwischen zwei Gruppen von Proben eingeführt. Unser Konzept basiert auf einer zirkulären binären Segmentierung, die mit Hilfe einer Scoring-Funktion Subregionen detektiert, die eine stärkere Differenz der mittleren Methylierungrate der beiden Gruppen zeigen, als der sie umgebende Hintergrund. Diese Subregionen werden mit dem zweidimensionalen Kolmogorov Smirnov-Test [Fasano *et al.*, 1987] auf signifikante Unterschiede getestet, wobei die Methylierungsraten aller Proben von beiden Gruppen berücksichtigt werden. Der Algorithmus detektiert DMRs sehr schnell und kann auch auf sehr großen Gruppen arbeiten. Vergleiche auf simulierten und realen Datensätzen zeigen, dass `metilene` andere bereits bestehende Methoden übertrifft und sich besonders gut eignet für verrauschte Datensätze wie sie beispielsweise häufig in der Krebsanalyse gefunden werden.

Im Rahmen dieser Arbeit werden die zuvor eingeführten Methoden und Algorithmen verwendet, um einen WGBS Datensatz bestehend aus Proben zweier verschiedener Subtypen von Keimzentrums-B-Zell-Lymphomen und gesunden Kontrollen zu analysieren. Unter Verwendung des zuvor vorgestellten Algorithmus wurden DMRs zwischen den drei Gruppen berechnet. Es wurde beobachtet, dass

DMRs unmittelbar abwärts von der Transkriptionsstartstelle stark angereichert sind, was auf eine regulatorische Relevanz dieser Regionen hinweist. Die Integration von Genexpressionsdaten zeigte, dass eine beträchtliche Menge der DMRs signifikante Korrelation zwischen Genexpression und DNA-Methylierung aufweist. Schlussendlich wurde die Information über Transkriptionsfaktor-Bindungsstellen und Mutationsdaten mit der Methylierung- und Expression-Datenanalyse kombiniert. Dadurch wurden Signalwege und Krebs-Subtyp spezifische Gene identifiziert, die stark verändert waren.

Schließlich wurden einige Erkenntnisse aus der Lymphomstudie über auf eine große Probenmenge, bestehend aus einer Vielzahl von Krebsarten, erweitert. Wir konnten zeigen, dass das Verhältnis der DNA-Methylierung von als Poised Promotor klassifizierten Regionen zu der Hintergrundmethylierungsrate ausreichend ist, um auf Basis von DNA-Methylierungsdaten von 450k Arrays fast alle Proben erfolgreich in Krebs- oder nicht-Krebs-Gewebe zu klassifizieren. Darüber hinaus haben wir festgestellt, dass in fast allen Krebs-Frischgewebe-Proben der Anstieg in der Methylierungsrate häufig mit der Hochregulation der Genexpression von Poised Promoter regulierten Genen zusammen fällt, was ein de-poising der Regionen impliziert.