Non-negative matrix factorization is applied to infer cellular composition and constituent gene expression programs from single-cell RNA-seq data

Inaugural-Dissertation
zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Universität zu Köln

vorgelegt von
Yasir Arafat Tamal
aus Tangail, Bangladesch

Köln, Oktober 2023

Abstract

Most cells in an organism contain the same genome, but the unique molecular instructions encrypted within the individual cells result in variability in cell-to-cell transcriptomics profile. These unique molecular instructions are sets of genes induced together as a gene expression program (GEP) through complex transcriptional co-regulation to establish a cell's identity (identity GEP) or to perform a cellular activity (activity GEP). These unique molecular fingerprints distinguish one individual cell from another. Single-cell RNA sequencing (scRNA-seq) technology enables us to capture variation in gene expression across many individual cells, which allows us to identify the constituent cell types and the gene expression programs (GEPs) that characterize these cell types for the biological tissue of interest. But due to the substantial drop-out effect and noisy nature of the high-dimensional scRNA-seq data, inferring the GEPs becomes difficult, necessitating analytical approaches to detect the underlying structure. In this thesis, I used a matrix factorization technique called non-negative matrix factorization (NMF) to infer the molecular makeup of cells by identifying identity GEPs and activity GEPs from scRNA-seq data and using this information to estimate the cellular makeup of *Cardamine hirsuta* and *Arabidopsis thaliana*. The unique ability of the NMF method to summarize scRNA-seq data by finding sets of coexpressed genes (GEPs) and representing cells by their GEPs usage makes it an attractive tool for scRNA-seq data analysis. The inferred GEPs by the NMF method reflect the signals present in the data by detecting biologically meaningful physiological activities or unwanted technical artifacts. It enables us to investigate the source of variation in each individual cell or cell type and gives us the freedom to keep the true molecular signals we are interested in. A common goal in developmental biology studies is understanding the genetic basis of phenotypic changes. Two developmental regulators, *STM* and *RCO*, are associated with the complex leaf phenotype of *C. hirsuta*. But the molecular network regulated by these two genes is still unexplored. In this study, I aimed to decipher the genetic basis underlying the divergent leaf form of *C. hirsuta* and *A. thaliana* using the NMF method. While reaching that goal, I established a cellular taxonomy of wild-type *C. hirsuta* leaf primordia by identifying and characterizing the cell types within the leaf tissue. I have defined two new cell types, the *RCO* and the *STM* cell types, which recapitulated the known biology of *RCO* and *STM* at the cellular level. The markers proposed for these two cell types can be the potential downstream targets of the two developmental regulators. In the search for the direct targets of *RCO*, this study proposed a set of 14 markers that were up-regulated in the WT genotype compared to the *rco*-mutant genotype of *C. hirsuta*. Finally, the comparative analysis between the two species, *A. thaliana*, and *C. hirsuta*, provided some key insights on the shared and species-specific features between the two species at the cellular level. A key finding of this comparative study, the identification of the *C. hirsuta*-specific STM cell type and a conserved RCO cell type between the species, reflected some of the key evolutionary developmental biology aspects of the evolution of the complex leaf morphology in *C. hirsuta*.

Zusammenfassung

Die meisten Zellen in einem Organismus enthalten dasselbe Genom, aber die einzigartigen molekularen Anweisungen, die in den einzelnen Zellen verschlüsselt sind, führen zu einer Variabilität des Transkriptomikprofils von Zelle zu Zelle. Bei diesen einzigartigen molekularen Anweisungen handelt es sich um Sätze von Genen, die zusammen als Genexpressionsprogramm (GEP) durch komplexe transkriptionelle Koregulierung induziert werden, um die Identität einer Zelle zu bestimmen (Identitäts-GEP) oder eine zelluläre Aktivität durchzuführen (Aktivitäts-GEP). Diese einzigartigen molekularen Fingerabdrücke unterscheiden eine einzelne Zelle von einer anderen. Die Technologie der Einzelzell-RNA-Sequenzierung (scRNA-seq) ermöglicht es uns, die Variation der Genexpression in vielen einzelnen Zellen zu erfassen. Dadurch können wir die einzelnen Zelltypen und die Genexpressionsprogramme (GEPs) identifizieren, die biologischen Gewebe charakterisieren. Aufgrund des erheblichen Drop-out-Effekts und der stochastischen Natur der hochdimensionalen scRNA-seq-Daten ist es jedoch schwierig, die GEPs abzuleiten, so dass analytische Ansätze erforderlich sind, um die zugrunde liegende Struktur zu erkennen. In dieser Arbeit habe ich eine Matrixfaktorisierungstechnik - nicht-negative Matrixfaktorisierung (NMF) - verwendet, um den molekularen Aufbau von Zellen abzuleiten. Dabei habe ich Identitäts-GEP und Aktivitäts-GEP aus scRNA-seq-Daten identifiziert und diese Informationen zur Schätzung des zellulären Aufbaus von *Cardamine hirsuta* und *Arabidopsis thaliana* verwendet. Die einzigartige Fähigkeit der NMF-Methode, scRNA-seq-Daten durch das Auffinden von Gruppen koexprimierter Gene (GEPs) zusammenzufassen und Zellen durch ihre GEPs-Nutzung darzustellen, macht sie zu einem attraktiven Werkzeug für die scRNA-seq-Datenanalyse. Die durch die NMF-Methode abgeleiteten GEPs spiegeln die in den Daten vorhandenen Signale wider, indem sie biologisch bedeutsame physiologische Aktivitäten oder unerwünschte technische Artefakte erkennen. Diese Methode ermöglicht es uns, die Quelle der Variation in jeder einzelnen Zelle oder jedem Zelltyp zu untersuchen, und gibt uns die Möglichkeit, diejenigen molekularen Signale zu identifizieren, an denen wir interessiert sind. Ein gemeinsames Ziel entwicklungsbiologischer Studien ist das Verständnis der genetischen Grundlage phänotypischer Veränderungen. Zwei Entwicklungsregulatoren, *STM* und *RCO*, werden mit dem komplexen Blattphänotyp von *C. hirsuta* in Verbindung gebracht. Das molekulare Netzwerk, das von diesen beiden Genen reguliert wird, ist jedoch noch unerforscht. Zweck dieser Studie war, mit Hilfe der NMF-Methode die genetischen Grundlagen für die unterschiedliche Blattform von *C. hirsuta* und *A. thaliana* zu entschlüsseln. Um dieses Ziel zu erreichen, habe ich eine zelluläre Taxonomie der Wildtyp-Blattprimordien von *C. hirsuta* erstellt, indem ich die Zelltypen im Blattgewebe identifiziert und charakterisiert habe. Ich habe zwei neue Zelltypen definiert, den RCO- und den STM-Zelltyp, die im Einklang sind mit der bekannten Biologie von RCO und STM auf zellulärer Ebene. Die für diese beiden Zelltypen vorgeschlagenen Markergene können die potenziellen nachgeschalteten Ziele der beiden Entwicklungsregulatoren sein. Auf der Suche nach den direkten

Zielen von *RCO* wurden in dieser Studie 14 Marker-gene vorgeschlagen, die im WT-Genotyp im Vergleich zum rco-mutierten Genotyp von *C. hirsuta* hochreguliert waren. Schließlich lieferte die vergleichende Analyse zwischen den beiden Arten *A. thaliana* und *C. hirsuta* einige wichtige Erkenntnisse über die gemeinsamen und artspezifischen Merkmale der beiden Arten auf zellulärer Ebene. Ein zentrales Ergebnis dieser vergleichenden Studie, die Identifizierung des *C. hirsuta*-spezifischen STM-Zelltyps und eines zwischen den Arten konservierten RCO-Zelltyps, spiegelt einige der wichtigsten entwicklungsbiologischen Aspekte der Evolution der komplexen Blattmorphologie von *C. hirsuta* wider.