Computational Methods for the Analysis of Gene Expression Regulation with Application to Hiv-1 Infection

Author(s):
Golumbeanu, Monica

Publication Date:
2017

Diss. ETH N° 24603

COMPUTATIONAL METHODS FOR THE ANALYSIS OF GENE EXPRESSION REGULATION WITH APPLICATION TO HIV-1 INFECTION

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZURICH

(Dr. sc. ETH Zurich)

presented by

MONICA GOLUMBEANU

Master of Science, Royal Institute of Technology (KTH) Stockholm
Master of Science (Technology), Aalto University Helsinki

born on 03. 07. 1988

citizen of Romania

accepted on the recommendation of

Prof. Dr. Niko Beerenwinkel, ETH Zurich
Prof. Dr. Renato Paro, ETH Zurich
Dr. Angela Ciuffi, University of Lausanne
Prof. Dr. Uwe Ohler, Max Delbrück Center Berlin

2017

Abstract

High throughput measurement techniques have provided significant advancement in tackling large-scale, challenging biological research questions, offering a systematic strategy to analyze complex biological systems and processes. They have been repeatedly utilized for identification and characterization of gene regulatory processes and gene expression modulation. The present thesis encompasses three individual computational analyses of high throughput gene expression data, generated on various cellular omics levels, which I conducted as a doctoral candidate in the Computational Biology Group at ETH Zürich. The three projects, organized in two parts within the current thesis, propose a diversified methodological palette aimed to study gene expression regulation.

RNA-binding proteins play a key role in regulating RNA throughout its lifespan, being responsible for a multitude of functions which still need to be fully elucidated. The first part of the present thesis focuses on RNA-protein regulatory interactions, precisely on the identification of genome-wide RNA-protein binding sites through analysis of PAR-CLIP data. Accordingly, Chapter 1 provides a synthesis of the biology and research in post-transcriptional regulation through RNA-binding proteins. Afterwards, Chapter 2 introduces BMix, a probabilistic method designed to extract RNA-protein binding sites from PAR-CLIP data. The method is based on a statistical model that explicitly accounts for the sources of noise in PAR-CLIP data to identify *bona fide* binding-induced T-to-C substitutions. The superior speed and accuracy of our method compared to existing approaches is demonstrated on both simulated, as well as publicly available, human datasets.

The second part of the thesis proposes two computational analyses characterizing regulatory programs during HIV-1 infection. The rapidly evolving complexity of HIV-1 pathogenicity and virulence has been obstructing the numerous attempts at a cure. Currently, antiretroviral drugs targeting various steps of the replication cycle keep the HIV-1 virus levels in infected patients to a low, manageable level. However, stopping antiretroviral therapy results in virus rebound, indicating the presence of reservoirs. Although still debated, the viral reservoirs could arise from circulating latently-infected cells or from productively infected cells residing in anatomical sanctuaries. Research efforts are still necessary in order to identify and characterize virus-host interactions during productive or latent infection. After a comprehensive overview of the current state in HIV-1 research done in the second part of Chapter 1, Chapter 3 exposes a research investigation aiming to define the dynamic proteo-transcriptomic response of T cells to HIV-1 infection. The chapter presents a statistical analysis that stratifies host cellular genes based on their expression temporal patterns at RNA, protein and phospho-protein levels. Next, Chapter 4 presents a study of cellular heterogeneity of HIV-1 latency and reactivation using the newly-developed single-cell RNA-Seq technology. Accordingly, an elaborate analysis methodology, capable of dealing with the high levels of noise exhibited by single-cell RNA-Seq data, reveals the transcriptional heterogeneity of HIV and host cell gene expression during latency and reactivation in T-cells.

The thesis concludes with a discussion of all the presented studies, their limitations and future perspectives, included in Chapter 5.

Résumé

Les techniques de mesure à haut débit ont apporté des progrès significatifs dans la poursuite des questions de recherche biologique à grande échelle, en offrant une stratégie systématique d'analyse des systèmes et processus biologiques complexes. Ces techniques ont été utilisées à plusieurs reprises pour l'identification et caractérisation des processus liés à la régulation des gènes et de la modulation de l'expression génique. Cette thèse propose une palette méthodologique diversifiée englobée en trois analyses computationnelles des données d'expression génique à haut débit, générées aux différents niveaux omiques, que j'ai menées en tant que doctorante au sein du Groupe de Biologie Computationnelle à ETH Zürich.

Les protéines de liaison à l'ARN jouent un rôle clé dans la régulation de l'ARN tout au long de sa vie, étant responsables d'une multitude de fonctions qui doivent encore être pleinement élucidées. La première partie de la thèse se concentre sur les interactions régulatrices entre l'ARN et les protéines, précisément sur l'identification à l'échelle génomique des sites de liaison des protéines à l'ARN, grâce à l'analyse des données PAR-CLIP. En conséquence, la première partie du Chapitre 1 fournit une synthèse de la biologie et de la recherche dans la régulation post-transcriptionnelle à travers des protéines de liaison à l'ARN. Ensuite, le Chapitre 2 présente BMix, une méthode conçue pour extraire des sites de liaison de protéines à l'ARN à partir de données PAR-CLIP. La méthode est basée sur un modèle statistique capable de prendre en compte les sources de bruit dans les données PAR-CLIP pour identifier les substitutions T-à-C indicatrices des liaisons RNA-protéine.

La deuxième partie de la thèse propose deux analyses computationnelles caractérisant les programmes de régulation génique au cours de l'infection cellulaire avec le virus VIH-1. La complexité et l'évolution rapide de la pathogénicité et de la virulence du VIH-1 a entravé des nombreuses tentatives de guérison. À l'heure actuelle, les médicaments antirétroviraux ciblant les différentes étapes du cycle de réplication virologique maintiennent les niveaux de VIH-1 chez les patients infectés à un niveau faible et tolérable. Cependant, l'arrêt de la thérapie antirétrovirale entraîne un rebond de virus, ce qui indique la présence des réservoirs viraux. Des efforts de recherche sont encore nécessaires pour identifier et caractériser les interactions virus-hôte pendant une infection productive ou latente. Après un aperçu complet de l'état actuel de la recherche sur le VIH-1, menée dans la deuxième partie du Chapitre 1, le Chapitre 3 expose une étude de recherche visant à décrire la réponse protéo-transcriptomique dynamique des cellules immunitaires humaines suite à leur infection par le VIH-1. Le chapitre présente une analyse statistique qui stratifie les gènes cellulaires en fonction de leur motifs d'expression temporelle aux niveaux de l'ARN, des protéines et des phospho-protéines. Ensuite, le Chapitre 4 présente une étude qui révèle l'hétérogénéité transcriptionnelle du VIH et de l'expression génique de la cellule hôte pendant la latence et la réactivation du VIH-1 à l'aide de la nouvelle technologie RNA-Seq appliquée aux cellules individuelles.

La thèse se termine par une discussion dans le Chapitre 5 sur les trois études présentées, leurs limites et leurs perspectives d'avenir.