TELOMERE ANALYSIS BASED ON HIGH-THROUGHPUT MULTI -*OMICS* DATA

Lilit Nersisyan

ABSTRACT

After their discovery, telomeres have been considered as the secret to longevity and cancer treatment. However, despite the excitement and the boom, methodologies to study telomeres are still low-throughput in nature, and do not take advantage of high-throughput data generation technologies, such as microarrays and next-generation sequencing technologies. These data also contain "undisclosed" information about telomeres and telomere related regulatory processes. The main aim of this thesis was to develop algorithms and software packages to use these high-throughput data in order to utilize the "hidden" information on telomere length dynamics and telomere length maintenance processes, and analyse telomere biology at a systems scale.

We have developed the software package Computel for estimation of mean telomere length from whole genome sequencing (WGS) data. Computel has the advantage of extracting telomere length information, aside from other genetic variations, from WGS data. We have applied it to real-world datasets to find associations between telomere length dynamics and transcriptomic/epigenomic features. An association study between genomic variations and telomere length in a healthy population of South Asians, led to identification of polymorphisms in *ADARB2* gene linked to longer telomeres. This finding and the previously reported link between *ADARB2* and extreme longevity make this gene a good candidate for future studies. In another study, we have utilized WGS, transcriptomic and epigenomic datasets on lung adenocarcinoma cell lines to mine the associative relationship between telomere length, gene expression, and epigenetic changes. We have identified several genes that might be in a regulatory relationship with telomeres.

In the second part of this thesis, we have applied a systems biology approach to study the two types of telomere length maintenance mechanisms (TMM): one based on the action of telomerase, the other based on homologous recombination events (ALT). There are many studies investigating the role of one or two molecular factors in these processes. However, to date, there are no unified

pathways that describe the TMMs. Additionally, the absence of properly annotated functional categories or pathways for TMMs makes standard systems biology analysis pipelines infeasible.

In this thesis, we have come up with an iterative algorithm to generate TMM pathways and to assess activity of TMM mechanisms from gene expression data. It is based on literature-based curation of TMM pathways, and their recursive validation – based on an in-house Pathway Signal Flow algorithm for pathway activity determination. We have validated our approach using a set of experimentally annotated cell lines and tumor cells. Our algorithm may serve as a convenient model to question the role of various molecular factors (proteins, RNAs, genetic mutations, etc) and interactions in TMMs, based on pathway extension and accuracy estimation.

In summary, we have developed a number of computational approaches to utilize high-throughput data for telomere research. We believe that these methodologies will help to utilize existing massive data in a more efficient manner for telomere-related studies and will foster telomere research. We have also presented a few of preliminary findings made using our software and algorithms.