Cell Tree Age as a new evolutionary model for representing age-associated somatic mutation burden

Booklet

Attila Csordas

Ph.D. Thesis

Szeged
2024

Cell Tree Age as a new evolutionary model for representing age-associated somatic mutation burden

Attila Csordas
Ph.D. Thesis

Appointed Consultant: Prof. Dr. Lajos Kemény

University of Szeged
Faculty of Medicine
Doctoral School of Clinical Medicine
Szeged
2024

Summary

Biological age is typically estimated using biomarkers whose states have been observed to correlate with chronological age. A persistent limitation of such aging clocks is that it is difficult to establish how the biomarker states are related to the mechanisms of aging. Somatic mutations could potentially form the basis for a more fundamental aging clock since the mutations are both markers and drivers of aging and have a natural timescale. Cell lineage trees inferred from these mutations reflect the somatic evolutionary process and thus, it has been conjectured, the aging status of the body. Such a timer has been impractical thus far, however, because detection of somatic variants in single cells presents a significant technological challenge.

Here we show that somatic mutations detected using single-cell RNA sequencing (scRNA-seq) from thousands of cells can be used to construct a cell lineage tree whose structure correlates with chronological age. De novo single-nucleotide variants (SNVs) are detected in human peripheral blood mononuclear cells using a modified

protocol. A default model based on penalized multiple regression of chronological age on 31 metrics characterizing the phylogenetic tree gives a Pearson correlation of 0.81 and a median absolute error of ~4 years between predicted and chronological age. Testing of the model on a public scRNA-seq dataset yields a Pearson correlation of 0.85. In addition, cell tree age predictions are found to be better predictors of certain clinical biomarkers than chronological age alone, for instance glucose, albumin levels and leukocyte count.

The geometry of the cell lineage tree records the structure of somatic evolution in the individual and represents a new modality of aging timer. In addition to providing a numerical estimate of 'Cell Tree Age', it unveils a temporal history of the aging process, revealing how clonal structure evolves over life span. Cell Tree Rings complements existing aging clocks and may help reduce the current uncertainty in the assessment of geroprotective trials.

Encoding accumulating somatic mutation burden in humans via evolutionary Cell Trees and representing the overall snapshot somatic mutation burden numerically as

a Cell Tree Age provides a new model to summarize the most fundamental hallmark of aging, using perhaps the most well established quantitative methodology in biology, phylogenetics.

Research objectives

The central conjecture behind our proposed aging timer is that the structure, or "shape", of cell trees is a representation of the biological aging process. There are two reasons for this hypothesis. The first is that phylogenetic systematics has long shown how genetic distances between species existing today reflect evolutionary changes in the past. It is reasonable then to expect that genetic distances between single cells can be used to infer the somatic evolutionary history of cells, a driver and indicator of aging. The second is that biomedical life history can leave its imprint on the cell tree, providing a record of major transitions in the aging process. An additional benefit of cell trees is that they provide an intuitively appealing representation of the

dynamics of aging that naturally lends itself to interpretation.

Using human peripheral blood cells from healthy individuals (n=18, age range 21-82 years of age) we have developed a new aging timer called Cell Tree Rings (CTR) with the following characteristics:

1. Naturally occurring somatic single nucleotide variants (SNVs) are used to build cell trees using standard phylogenetic algorithms,
2. SNVs are called directly and de novo from scRNA-seq data from hundreds or thousands of cells,
3. A broad set of tree metrics is used to identify aspects of tree shape that are associated with chronological age using a penalized multiple regression model,
4. The model is used to predict a Cell Tree Age for individuals.

Two different types of phylogenetic algorithms, UPGMA and maximum likelihood, are shown to produce a working Cell Tree Age model, providing extra evidence for the

hypothesis. Importantly, the model is validated with public data as an independent test set. The predicted Cell Tree Ages are also shown to correlate with some clinical blood biomarkers, for instance glucose, albumin, leukocytes, and monocytes.

Methods

Experimental data and protocol

Biological sample collection and isolation of cells: 18 blood samples, 5 ml each, have been collected by venipuncture at the Healthy Longevity Clinic (HLC) in Prague, Czech Republic. The samples have been taken with informed consent from healthy patients of the clinic. The Healthy Longevity Clinic Ethical Committee has reviewed and approved the Tree Ring Pilot observational study protocol with the reference number 20220301_001. The age range of the volunteers was 21-82 years old at the time of blood collection, 10 volunteers were males and 8 were females. Samples were processed by the same protocol. Viable peripheral blood mononuclear cells (PBMCs) were isolated from the collected biological sample.

The cells were labelled with molecular tags or CellPlex. Cells were loaded on the Chromium Controller and libraries prepared according to the original protocol CG000390 Chromium Next GEM Single Cell3 v3.1 Cell Surface Protein Cell Multiplexing RevB, aiming for 16000 recovered cells. Some library preparation steps were modified slightly. During cDNA amplification, the polymerization step was extended to 1.5 min. After cDNA purification, the samples were split into two aliquots (A and B) that were processed in parallel and differed only in the implementation of size selection.

After PCR amplification, both samples were purified using SPRIselect beads.

Sequencing: Library pools were sequenced on an Illumina NovaSeq 6000 using the S4 300-cycle kit and with 150 bp long R2.

General schematics of the Cell Tree Rings computational workflow

The Cell Tree Rings computational pipeline involves four consecutive steps, names of the sub-pipelines highlighted in bold.

1. **Tizkit**: Barcode specific calling of SNVs with scSNV and germline filtering.
2. **Tiznit**: Generating fasta files and phylogenetic inference of cell trees.
3. **AgeTreeShape:** Compute tree features and univariate regression on age.
4. **CellTreeAge:** Building multiple penalised regression model and Cell Tree Age prediction.

Results

Cell tree metrics

The central hypothesis of the study is that the shape of cell trees is a measure of biological age. Here shape refers to the combination of topology (branching order) and branch lengths. Topology, in the case of cell lineage trees, corresponds to branching patterns of mitotic division in somatic cells. Branch lengths represent the amount of evolutionary change and are usually defined as the product of mutation rates and a suitable unit of time.

Different tree metrics capture either topology-only, branch-length-only, or a combination of both. We have

applied a set of 31 tree metrics to characterize the shape of cell trees built from somatic mutations from human peripheral blood mononuclear cells. This set of measures comprises both traditional tree metrics used in phylogenetics and some that were developed specifically for this study.

The tree metrics, or features, can be split into 5 groups based on their technical properties. Group I contains spectral tree metrics that are based on the transform matrices of the cell trees which are discrete analogues of the generalized Fourier and Laplacian transforms, respectively. Group II contains specialised phylogenetic features focusing on aggregated branch length statistics and their derivatives, including entropy-based metrics. Group III includes well-known general phylogenetic tree statistics used in the biodiversity literature. Group IV focuses on branch length values specifically and generates summary statistics based on the distance matrix between the tips of the tree. Finally, Group V has 2 powergraph based features generating the Laplacian transforms of the square of the tree graphs, similar to Group I.

Default model building and prediction results with all tree features

An age predictor has been constructed by regressing chronological age on the 31 tree metrics, along with Sex. Elastic Net regularization and nested leave-one-out cross validation were used to validate and test the model. The resulting prediction errors, correlation coefficient, p-values and explained variance were estimated by comparing Cell Tree Ages to chronological ages.

In the default model, UPGMA is used for phylogenetic tree inference, 5 pseudo-replicate trees are used per sample, and all 32 features are used in the regression. For this model, r=0.813, p=0.00004, R^2=0.660, MAE=7.625, MdAE=4.396, and RMSE=10.760.

Public Validation of Cell Tree Age Model

Having developed the Cell Tree Age Model on HLC data, we then tested it on a public scRNA-seq dataset. For this we used 18 peripheral blood samples from the Asian Immune Diversity Atlas (AIDA). This involved 10

females and 8 males ranging in age between 21 – 65 years old. For these data, as with the HLC data, we used UPGMA to generate 5 pseudo-replicate trees from each sample where each pseudo-replicate used 700 cells randomly selected from the sample. The default model was trained on all 18 of the HLC data. The resulting performance metrics were r=0.853, p = 0.00001, R^2=0.728, MAE=12.791, MdAE=13.636, RMSE=14.081.

Cell Tree Age associations with clinical biomarkers

As part of the approved study protocol, a wide range of clinical blood biomarker work has been shared. These were obtained during the same period as the blood draws for the scRNA-seq. For the following list of clinical blood markers, we had access to values of 13 samples out of the 18 HLC samples, 9 males and 4 females: complete leukocyte and erythrocyte blood panel (13 markers), albumin, glucose, hba1c, total cholesterol, alkaline phosphatase, uric acid, creatinine, blood urea nitrogen. This provided a total of 21 markers.

Out of the 21 blood markers, 8 markers were associated significantly with Cell Tree Age or Chronological Age or both. Six out of these 8 significant associations showed Cell Tree Ages to be stronger predictors than Chronological Age with the metabolic marker blood glucose showing the biggest difference.

LIST OF PUBLICATIONS

Scientific papers included in this thesis

I. **Csordas A**, Sipos B, Kurucova T, Volfova A, Zamola F, Tichy B, Hicks DG. Cell Tree Rings: the structure of somatic evolution as a human aging timer. *Geroscience*. 2024 Jan 4. doi: 10.1007/s11357-023-01053-4. Epub ahead of print. PMID: 38172489. **Q1 with IF:5.7**

II. **Csordas A**, Cell lineage trees: the central structure plus key dynamics of biological aging and formulating the limiting problem of comprehensive organismal rejuvenation. *PeerJ Preprints* (2019). **Perspective, not peer reviewed. https://peerj.com/preprints/27821v7/**