A data- and model-driven analysis reveals the multi-omic landscape of ageing

Elisabeth Yaneske and Claudio Angione

Department of Computer Science and Information Systems, Teesside University, UK

Keywords: ageing, Metabolomic age, Transcriptomic age, Flux balance analysis, Multi-omics, CD4 T-cells.

Abstract

Altered expression of a number of genes has been correlated to biological ageing in humans. The biological age predicted from gene expression levels is known as transcriptomic age. This differs from chronological age which is measured as the time that an individual has lived since their date of birth. Transcriptomic age can be older or younger than an individual's chronological age. At present, studies have focused on using transcriptomic data to predict transcriptomic age. However, this approach largely does not consider the effect that genes have on the metabolic network and therefore on the observable cellular phenotype. This research takes the current understanding of transcriptomic ageing a step further by generating and investigating genome-scale metabolic models of ageing, using machine learning methods and a multi-omic approach based on constraint-based modelling. We combine these models with a transcriptomic age predictor and gene expression data from CD4 T-Cells from human peripheral blood mononuclear cells in healthy individuals. We show that metabolic models augmented with transcriptomics data of ageing can generate greater metabolic insights into the differences between chronological and transcriptomic age. Compared to standard transcriptomic-only approaches, our method provides a more comprehensive analysis of transcriptomic ageing and paves the way for a multi-omic understanding of ageing mechanisms in human cells.

Introduction

Ageing is associated with phenotypes such as wrinkles, greying hair, hair loss and frailty as well as diseases such as diabetes, osteoporosis and cardiovascular disease. Underlying these characteristics of ageing are age-related changes to our metabolic processes [1]. Changes to mitochondrial function, which is vital for energy metabolism and homeostasis, have been linked to the ageing process [2, 3]. Metabolic dysregulation leads to the build-up of fat stores in the abdominal cavity [4]. This type of fat is called visceral fat. Higher visceral fat levels mean a greater risk of insulin resistance, type 2 diabetes, cardiovascular disease and

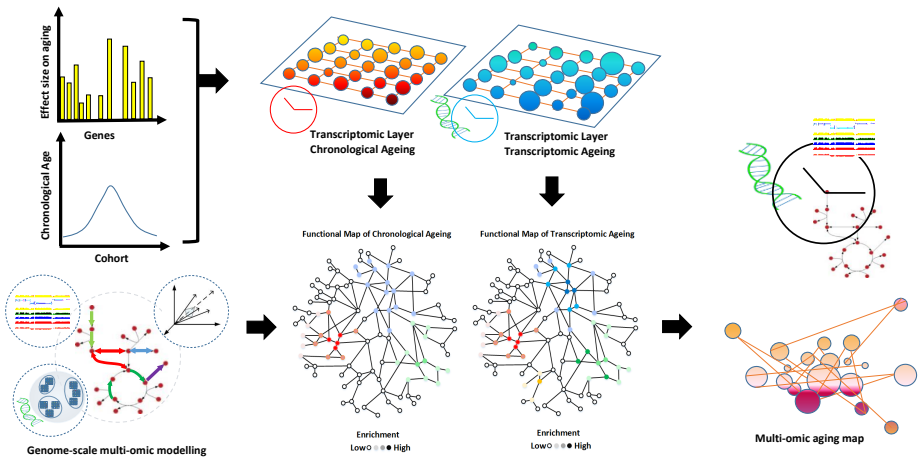

Fig. 1. Multi-omic ageing pipeline. We use the chronological data and corresponding predictors to obtain the effect of both types of ageing on the transcriptomic layer. We then combine the latter with the functional biological network data determined by the metabolism and multi-omic model to obtain the multi-omic ageing map.

cancer [5–8]. However, a recent study by Chee et al. [9] suggests that some age-related metabolic changes may not be inevitable but rather due to lifestyle. A group of young (mean age 21.5) and older (mean age 69.7) participants who were matched for physical activity levels and body composition were found to have no significant difference in insulin resistance and lipid accumulation in their muscles.

The term 'inflammaging' was proposed by Franceschi et al. [10] to describe the imbalance between inflammatory and anti-inflammatory networks which contribute to the chronic diseases of ageing. The immune response declines dramatically with ageing leading to increased frailty e.g. susceptibility to pathogens such as influenza [11]. CD4 cells have recently been shown to be linked to the ageing process. A study comparing gene expression from the same CD4 cells in newborns, middle-aged and long-lived participants found significant expression changes in the transcriptomes across the age groups [12]. The action of CD4 cells is impaired with age which contributes to the deterioration of the immune response [13, 14].

Although genes are often regarded as the main players in deciding cell behaviour, it is difficult to predict how they affect the phenotype. Metabolic models can therefore be used to predict the metabotype of a cell, and provide a close-to-phenotype prediction of cell behaviour and fate. Furthermore, metabolic models can be constrained using gene expression data [15].

The idea of this paper is that modelling how the gene expression alterations change metabolic processes provides greater understanding of the ageing process and a more accurate prediction of biological age. This data can

then be used to investigate how adding the metabolic information alters and improves transcriptomics-only classifications of patients. Using transcriptomic data from patients, we here generate patient-specific genome-scale metabolic models. We then investigate such models in the context of ageing, therefore obtaining metabolic ageing biomarkers. More specifically, the gene expression data from CD4 cells is mapped to both the chronological age and the transcriptomic age of the individuals in the cohort. This provided us with a baseline of the current understanding of transcriptomic ageing in the transcriptomic layer, and allowed us to investigate the metabolic processes that contribute to both chronological and transcriptomic ageing.

Methods

Transcriptomic age predictor

Throughout our pipeline, we use a meta-analysis of CD4 T-Cell gene expression from human peripheral blood mononuclear cells isolated from healthy individuals in the Boston area. The CD4 cell gene expression data was profiled on Affymetrix Human Gene 1.0 ST microarrays [16]. The raw data is accessible through GEO accession number GSE56033. The dataset contains 499 individual patients. Using 1497 age-associated genes, i.e. those found to be differentially expressed with chronological age, the effect size of each individual age-associated gene expression level on chronological age was then calculated [17]. The sum of the effect sizes of the age-associated genes on chronological age was used to calculate a general transcriptomic predictor Z, defined as a linear combination:

$$Z = \sum_i b_i x_i, \tag{1}$$

where x_i is the gene expression level of the ith probe, and b_i is the effect size for the ith probe.

Starting from these predictors, we calculated the transcriptomic age of each individual. To this end, the general transcriptomic predictor was scaled using the mean and standard deviation of the chronological age and the mean and standard deviation of the general predictor [17]. The transcriptomic age of an individual is:

$$SZ = \mu_{age} + (Z - \mu_z)\frac{\sigma_{age}}{\sigma_z}, \tag{2}$$

where μ_{age} and σ_{age} are the mean and the standard deviation of the chronological age, while μ_z and σ_z are the mean and the standard deviation of the predictor Z.

Using (2), we calculated the transcriptomic ages for the 499 individuals. The chronological data was included for each sample. The data was normalised using RMA (Robust Multi-array Average) [18] followed by dividing the gene expression values by the mean value for each probe. The effect size b_i is defined on a gene by gene basis. In cases where a probe represented more than one gene the effect

sizes for those genes were averaged to give an on overall average effect size for that probe.

To progress to a multi-omic understanding of transcriptomic ageing we used a constraint-based model of the CD4 cell [19] augmented with transcriptomics through METRADE [20]. This enabled us to create personalised metabolic models of the samples within the cohort. We then mapped the metabolic flux rate to the chronological age of the individuals within the cohort and repeated this pipeline using metabolic flux rate and transcriptomic age.

Constraint-based modelling

Given the matrix S of all known metabolic biochemical reactions and their stoichiometry, and given the vector v of flux rates in a given growth condition [21], constraint-based modelling and flux balance analysis (FBA) allow the prediction of the distribution of flux rates in a given condition. The metabolic network is solved by maximising a cellular objective (usually the biomass), with constraints deriving from the steady-state condition $Sv = 0$ and additional constraints $v^{\min}$ and $v^{\max}$ on lower bound and upper bounds of v.

Using METRADE [20] coupled with the transcriptomic data from each patient, we modify the upper- and lower- limits of reactions as a function of the expression levels of genes involved in the reaction. For each patient, to predict the cellular flux distribution when multiple objectives have to be taken into account, we use the following bilevel linear program:

$$\begin{aligned} &\max\ g^{\intercal}v \\ &\text{such that} \qquad \max\ f^{\intercal}v, \quad Sv = 0, \\ &\qquad\qquad v^{\min}\varphi(\Theta) \leq v \leq v^{\max}\varphi(\Theta). \end{aligned} \tag{3}$$

The vectors f and g are weights to select (or combine) the objectives to be maximised from the vector v. The vector Θ converts the gene expression values into coefficients for the bounds of reactions activated by those genes. This is achieved through the function φ, which acts on Θ, the expression of a biochemical reaction, is defined from the patient-specific expression levels of its genes, with a rule depending on the type of enzyme (single gene, isozyme, or enzymatic complex).

Cluster analysis

In order to compare the transcriptomic and metabolomic landscapes of ageing, Agglomerative Hierarchical Clustering (AHC) was used to cluster both the transcriptomic and fluxomic data. The transcriptomic data was RMA normalised followed by dividing the gene expression values by the mean value for each probe prior to clustering. AHC requires the distance between every pair of objects in the dataset to be calculated prior to being able link objects together into an hierarchical cluster tree based on proximity. AHC dissimilarity testing was

carried out to find the distance and linkage combination that gave the highest Cophenetic Correlation Coefficient (CHC). For the transcriptomic data, the linkage method that performed best was Average followed by Weighted, Complete, Ward then Single. Further testing showed that Euclidean distance and Average linkage gave the best CHC of 0.67. For the fluxomic data the best performing linkage method was Average followed by Single, Weighted, Ward then Complete. The highest CHC of 0.54 was obtained using Average linkage and Squared Euclidean distance.

The average linkage method used the Unweighted Pair-Group Method with Arithmetic Mean (UPGMA) [22] algorithm. UPGMA joins the pair of objects with the smallest distance, then calculates the average between this pair and all the other objects in the dataset, and repeats this cycle until all data has been grouped into one cluster. The distance is computed as:

$$D_{(ij),k} = \left(\frac{n_i}{n_i + n_j}\right) D_{ik} + \left(\frac{n_i}{n_i + n_j}\right) D_{jk}, \tag{4}$$

where D_{ij} is group (ij) which has $n(ij) = n_i + n_j$ members and k is a new cluster.

Results

In order to determine the optimal number of natural clusters the pairwise distance data for both the transcriptomic and fluxomic data were visualised using scatter plots (Figure 2). The most distinct categorisation occurred at four clusters for transcriptomic data. Figure 2 clearly shows the majority of the transcriptomic data forming two large clusters. For the fluxomic data the most distinct categorisation occurred at eight clusters. Figure 2 shows the fluxomic data falling into four main clusters, though less distinct than the transcriptomic data.

Having determined the optimal number of natural clusters for both transcriptomic and fluxomic data the data was further analysed to determine whether the natural clusters correlated with chronological age, transcriptomic age, or both. This analysis builds on recent research which has shown a correlation between CD4 cells and ageing [23]. Scatter plots of transcriptomic age against cluster number were produced for both transcriptomic and fluxomic data. These plots were annotated by chronological age using both colour and size. The results of the clustering with age are displayed in Figure 3 for the transcriptomic data and for the fluxomic data.

From Figure 3 it can be seen that although the transcriptomic data formed into two distinct clusters there is very little differentiation in chronological age between the two clusters. There is, however, some differentiation in the transcriptomic age with cluster one containing most of the older patients (approximately 38 and over) and cluster two containing most of the younger patients (approximately 29 and under). These results support the assertion that gene expression levels are related more to biological ageing than chronological ageing. However,

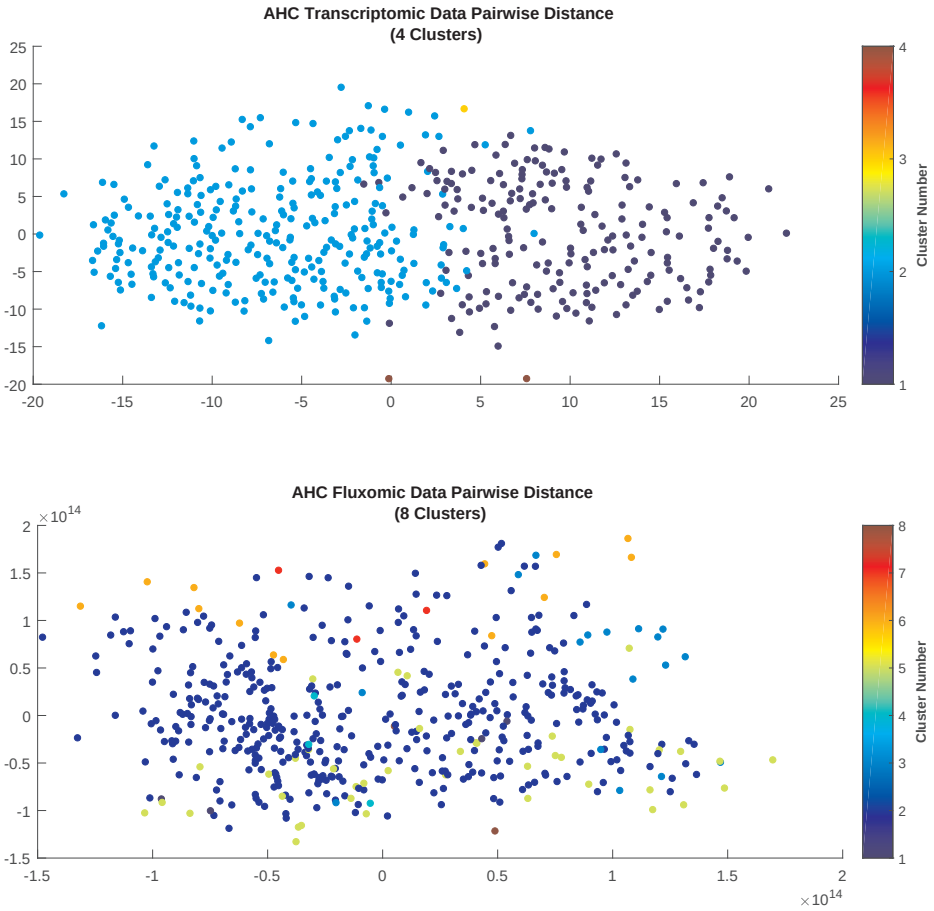

Fig. 2. The AHC pairwise distance matrix data. The pairwise distance plots were obtained after multidimensional scaling. (Top) The top scatter plot shows the pairwise Euclidean distance matrix for the transcriptomic data when classified into four clusters. Here we can see the majority of the data has formed into two distinct clusters. (Bottom) The bottom scatter plot shows the pairwise Squared Euclidean distance matrix for the fluxomic data when classified into eight clusters. The fluxomic data shows a more complex picture with most of the data forming into four less distinct clusters.

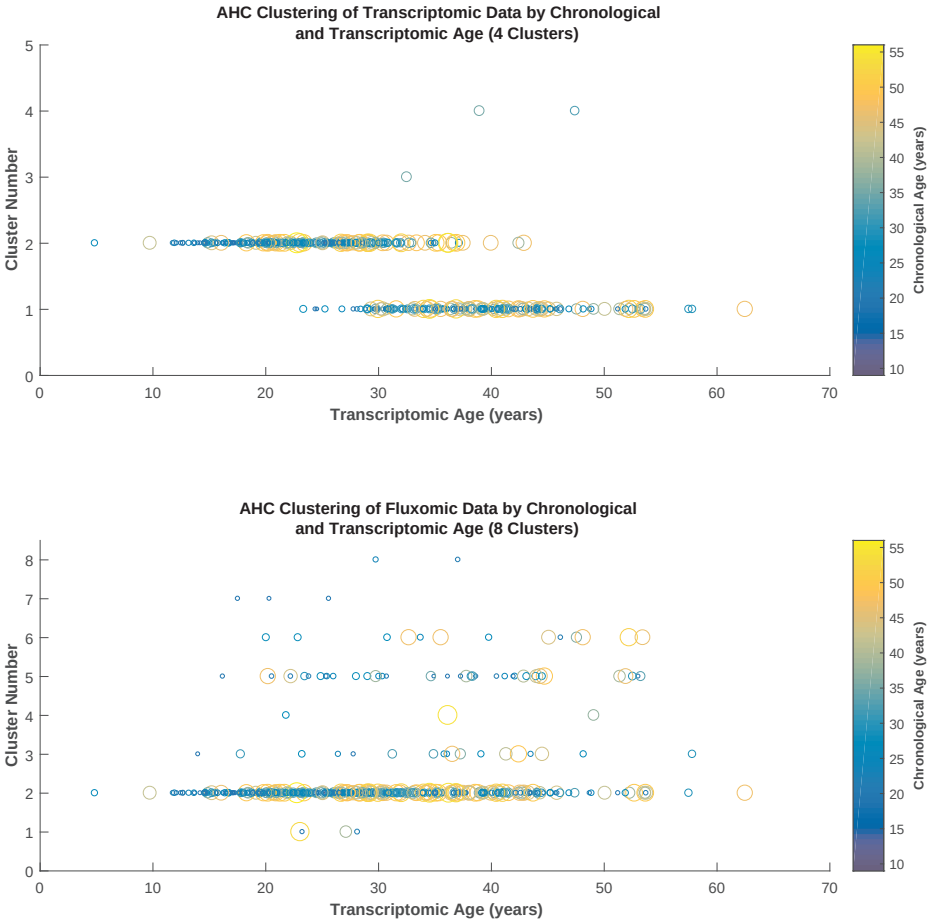

Fig. 3. AHC clustering of transcriptomic and fluxomic data with transcriptomic age annotated by chronological age. Chronological age is annotated using both colour and size. (Top) The top scatter plot shows clustering of the transcriptomic dataset. Cluster 1 and cluster 2 show some differentiation with transcriptomic age. Cluster 1 contains a greater proportion of the older transcriptomic age range from approximately 38 years and older. Cluster 2 contains most of the patients under 29. Very little differentiation can be seen according to chronological age. (Bottom) The bottom scatter plot shows clustering of fluxomic data. Although most of the data is captured in cluster 2, the fluxomic data shows some differentiation in both chronological and transcriptomic age. For chronological age, cluster 3 captures mainly the 20-30 year age range while cluster 5 appears to capture 30s to early 40s. For transcriptomic age, cluster 3 captures 30s to mid 40s, cluster 5 captures 20s, late 30s and early 40s and cluster 6 captures early 30s and late 40s. Cluster 7 and 8 contain small subsets of patients in the 20s and 30s respectively.

the clustering has a low level of granularity with not much differentiation between age groups.

The fluxomic data in Figure 3 shows a more complicated picture. Most of the data is contained in cluster two for both chronological and transcriptomic age. There is some differentiation in chronological ageing with cluster three appearing to capture mainly the 20-30 year age range and cluster five appearing to capture 30s to early 40s. There is also some differentiation in transcriptomic ageing with cluster 3 capturing 30s to mid 40s, cluster five capturing late 30s and early 40s and cluster six capturing early 30s and late 40s. The last two clusters (seven and eight) contain non-overlapping subsets of patients in their 20s and 30s respectively. This appears to show greater differentiation than the transcriptomic data between both chronological and transcriptomic ageing which suggests that further analysis could provide deeper insights into the metabolism of ageing.

The standard deviation was calculated to determine the within cluster variation of the chronological ages. Fluxomic data had the lowest average standard deviation overall of 5.6776 for 4 clusters. The lowest average standard deviation for the transcriptomic data was 6.0498 for 2 clusters. To determine the between cluster variation an F-test was performed. The fluxomic results for 4 clusters showed the difference of standard deviation between cluster 7 and the other clusters was statistically significant ($p < 0.05$). No significant results were found for transcriptomic data. These results show that, for this dataset, clustering by fluxomic improves the clustering by transcriptomics. Using a multi-variate method such as Principal Component Analysis for further investigation could allow the identification of the principal metabolic fluxes involved in different age groups and hence in the ageing process.

Conclusion

As with many diseases and health conditions, transcriptomics only is unlikely to reveal the full picture of ageing [24–26]. Multi-omic stoichiometric modelling can be used to predict the effects of gene expression on the metabolism and vice versa. More specifically, where a set of genes are known to be involved in a disease, genome-scale models can be integrated with transcriptomic data. Using a multi-omic model of ageing, we here achieve a greater understanding of the mechanisms of the disease by modelling the effect of altered gene expression on the metabolism of an organism (Figure 1).

Interestingly, we are able to obtain a metabolomic network of patients, as opposed to the transcriptomic-only network commonly obtained from gene expression data. We applied agglomerative hierarchical clustering to visualize functional mapping of local enrichment of the metabolic flux rate response to the chronological and transcriptomic ageing processes. We report that the correlation between the chronological age and transcriptomic age shows the relationship between the transcriptomic layer and the metabolomic layer.

Future analyses, such as principal component analysis of the multi-omic data will enable us to show which metabolic pathways correlate with ageing. We

will therefore be able to identify biomarkers that, taken together, can define a *metabolomic age predictor*. This can, for example, allow the classification of patients into different phenotypes depending on the metabolic presentation of the disease. Our model could also be used as a diagnostic/prognostic tool e.g. for early chronological ageing.

We finally showed that, using machine learning approaches, we can classify patients according to their metabolic phenotype of ageing by detecting global similarity between patients across multiple omics. As a result, we reach a multi-omic and genome-scale viewpoint that significantly extends the state-of-the-art, but transcriptomic-only, understanding of ageing.

References

1. Nir Barzilai, Derek M Huffman, Radhika H Muzumdar, and Andrzej Bartke. The critical role of metabolic pathways in aging. *Diabetes*, 61(6):1315–1322, 2012.
2. Ivana Bratic and Aleksandra Trifunovic. Mitochondrial energy metabolism and ageing. *Biochimica et Biophysica Acta (BBA)-Bioenergetics*, 1797(6):961–967, 2010.
3. Denham Harman. The biologic clock: the mitochondria? *Journal of the American Geriatrics Society*, 20(4):145–147, 1972.
4. William Cameron Chumlea, Robert L Rhyne, Philip J Garry, and William C Hunt. Changes in anthropometric indices of body composition with age in a healthy elderly population. *American journal of human biology*, 1(4):457–462, 1989.
5. Ike S Okosun, KM Chandra, Simon Choi, Jacqueline Christman, GE Dever, and T Elaine Prewitt. Hypertension and type 2 diabetes comorbidity in adults in the united states: risk of overall and regional adiposity. *Obesity*, 9(1):1–9, 2001.
6. Aaron R Folsom, Susan A Kaye, Thomas A Sellers, Ching-Ping Hong, James R Cerhan, John D Potter, and Ronald J Prineas. Body fat distribution and 5-year risk of death in older women. *Jama*, 269(4):483–487, 1993.
7. Radhika Muzumdar, David B Allison, Derek M Huffman, Xiaohui Ma, Gil Atzmon, Francine H Einstein, Sigal Fishman, Aruna D Poduval, Theresa McVei, Scott W Keith, et al. Visceral adipose tissue modulates mammalian longevity. *Aging cell*, 7(3):438–440, 2008.
8. Eugenia E Calle, Carmen Rodriguez, Kimberly Walker-Thurmond, and Michael J Thun. Overweight, obesity, and mortality from cancer in a prospectively studied cohort of us adults. *New England Journal of Medicine*, 348(17):1625–1638, 2003.
9. Carolyn Chee, Chris E Shannon, Aisling Burns, Anna L Selby, Daniel Wilkinson, Kenneth Smith, Paul L Greenhaff, and Francis B Stephens. Relative contribution of intramyocellular lipid to whole-body fat oxidation is reduced with age but subsarcolemmal lipid accumulation and insulin resistance are only associated with overweight individuals. *Diabetes*, 65(4):840–850, 2016.
10. Claudio Franceschi, Miriam Capri, Daniela Monti, Sergio Giunta, Fabiola Olivieri, Federica Sevini, Maria Panagiota Panourgia, Laura Invidia, Laura Celani, Maria Scurti, et al. Inflammaging and anti-inflammaging: a systemic perspective on aging and longevity emerged from studies in humans. *Mechanisms of ageing and development*, 128(1):92–105, 2007.
11. Bonnie B Blomberg and Daniela Frasca. Quantity, not quality, of antibody response decreased in the elderly. *The Journal of clinical investigation*, 121(8):2981–2983, 2011.

12. Ming Zhao, Jian Qin, Hanqi Yin, Yixin Tan, Wei Liao, Qian Liu, Shuangyan Luo, Min He, Gongping Liang, Yajing Shi, et al. Distinct epigenomes in cd4+ t cells of newborns, middle-ages and centenarians. *Scientific Reports*, 6:38411, 2016.
13. Moro-García Marco and Alonso-Arias Rebeca. When aging reaches cd4+ t-cells: Phenotypic and functional changes. *Frontiers in Immunology*, 4:107, 2013.
14. Julie S Lefebvre and Laura Haynes. Aging of the cd4 t cell compartment. *Open Longevity Science*, 6:83, 2012.
15. Claudio Angione, Max Conway, and Pietro Lió. Multiplex methods provide effective integration of multi-omic data in genome-scale models. *BMC bioinformatics*, 17(4):83, 2016.
16. Towfique Raj, Katie Rothamel, Sara Mostafavi, Chun Ye, Mark N Lee, Joseph M Replogle, Ting Feng, Michelle Lee, Natasha Asinovski, Irene Frohlich, et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science*, 344(6183):519–523, 2014.
17. Marjolein J Peters, Roby Joehanes, Luke C Pilling, Claudia Schurmann, Karen N Conneely, Joseph Powell, Eva Reinmaa, George L Sutphin, Alexandra Zhernakova, Katharina Schramm, et al. The transcriptional landscape of age in human peripheral blood. *Nature communications*, 6, 2015.
18. Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2):249–264, 2003.
19. Feifei Han, Gonghua Li, Shaoxing Dai, and Jingfei Huang. Genome-wide metabolic model to improve understanding of cd4+ t cell metabolism, immunometabolism and application in drug design. *Molecular BioSystems*, 12(2):431–443, 2016.
20. Claudio Angione and Pietro Lió. Predictive analytics of environmental adaptability in multi-omic network models. *Scientific reports*, 5, 2015.
21. Bernhard Ø Palsson. *Systems Biology: Constraint-Based Reconstruction and Analysis*. Cambridge University Press, 2015.
22. Charles D Michener and Robert R Sokal. A quantitative approach to a problem in classification. *Evolution*, pages 130–162, 1957.
23. Marie-Paule Vasson, Marie-Chantal Farges, Nicolas Goncalves-Mendes, Jérémie Talvas, Josep Ribalta, Brigitte Winklhofer-Roob, Edmond Rock, and Adrien Rossary. Does aging affect the immune status? a comparative analysis in 300 healthy volunteers from france, austria and spain. *Immunity & Ageing*, 10(1):38, 2013.
24. Mamas Mamas, Warwick B Dunn, Ludwig Neyses, and Royston Goodacre. The role of metabolites and metabolomics in clinically applicable biomarkers of disease. *Archives of toxicology*, 85(1):5–17, 2011.
25. Pranov Ramana, Erwin Adams, Patrick Augustijns, and Ann Van Schepdael. Metabonomics and drug development. *Metabonomics: Methods and Protocols*, pages 195–207, 2015.
26. Daniel W Nebert and Elliot S Vesell. Can personalized drug therapy be achieved? a closer look at pharmaco-metabonomics. *Trends in pharmacological sciences*, 27(11):580–586, 2006.