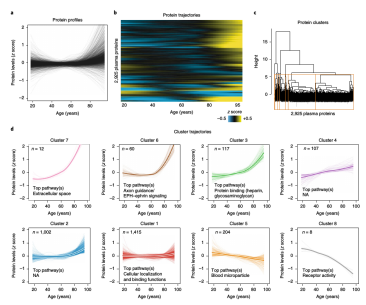
Кластеризация возрастных трендов экспрессии генов PBMC человека

Алексей Алексеев,
МГУ им. М.В. Ломоносова, физический факультет
alekseev@physics.msu.ru vk.com/aleksey3a

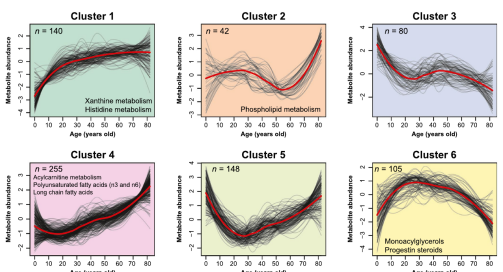
Введение

В настоящее время существует множество теорий старения человека, при этом для проверки этих теорий и построения системно-биологических моделей старения человека не хватает исходных данных в виде трендов, то есть изменения средних по популяции значений в омиксных данных [1]. В ряде недавних работ получены тренды по концентрации белков [2] и метаболитов [3] в крови, однако в настоящее время не развиты подходы для получения трендов экспрессии генов для клеток человека. Эта задача осложнена отсутствием в открытых источниках значимого объема данных экспрессии мРНК по различным тканям человека, а также межиндивидуальной и межполовой вариабельностью в этих данных. Как мы ожидаем, получение трендов экспрессии позволит сразу приобрести значительный материал для изучения изменений в сигнальных и метаболических путях при старении человека.

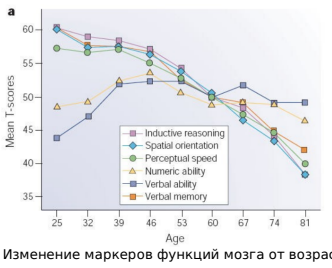
Примеры среднепопуляционных трендов, в том числе с кластеризацией



Протеомные данные из работы [1], с кластеризацией трендов



Метаболомные данные из работы [2], с кластеризацией трендов



Изменение маркеров функций мозга от возраста

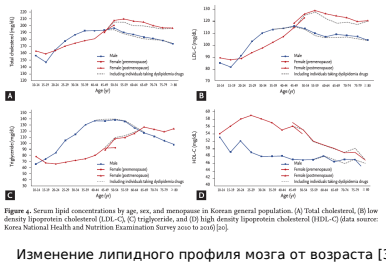
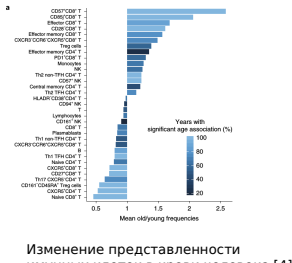


Figure 4. Serum lipid concentrations by age, sex, and menopause in Korean general population. (A) Total cholesterol, (B) low density lipoprotein cholesterol (LDL-C), (C) triglyceride, and (D) high density lipoprotein cholesterol (HDL-C) (data source: Korea National Health and Nutrition Examination Survey 2010 to 2016) [20].

Изменение липидного профиля мозга от возраста [3]



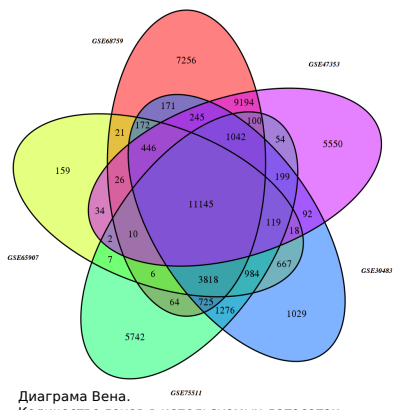
Изменение представленности иммунных клеток в крови человека [4]

Цели работы

- Разработать способ получения связанных со старением трендов экспрессионных данных
- Осуществить валидацию результатов, в том числе с использованием данных других работ
- Разработать подход к кластеризации трендов генов для развития методов поиска групп генов, имеющих сходное (синхронное) изменение в процессе старения.

Исходные данные

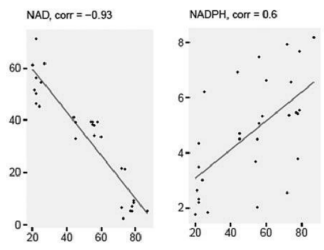
Мы воспользовались открытыми данными из базы GEO: 4 датасета для PBMC человека GSE75511, GSE30483, GSE47353, GSE68759 и опорный датасет GSE65907. Изучены гены, которые пересекаются по всем датасетам (11145 генов). Для образцов в каждом датасете указан пол и возраст. Из них отдельно были проанализированы 460 генов из метаболических путей, связанных с энергетикой клетки, TCA, путями регуляции инсулина и лептина, окислением жирных кислот в пероксисомах и митохондриях, а также все известные реакции с участием в NAD+ и NADPH.



Диаграма Вена.
Количество генов в используемых датасетах

group	count
AMPK-subunits	7
CDF-related	26
Fructose-methabolysm	10
Glutatione-metabolism	37
Glycose-related	62
Insuline-related	45
Leptin-related	26
NAD-related	258
NADPH-related	198
Peroxysome-related	83
Polyol-pathway	4
PPP-related	33
TCA-related	31

Таблица групп генов (всего 653), интересных с точки зрения изучения изменения энергетики клетки при старении — на основе базы KEGG и «ручного» обзора литературы

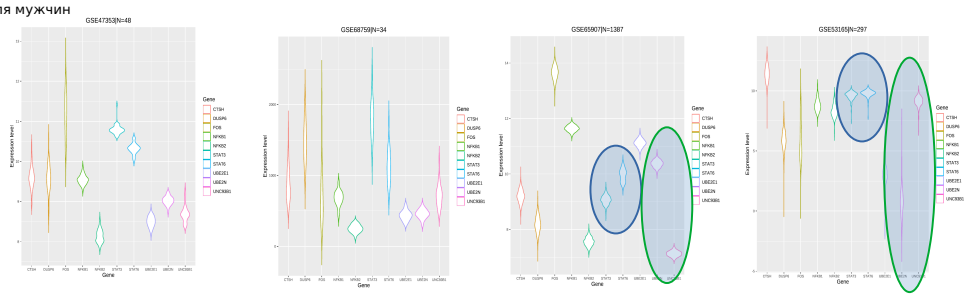


Изменение концентрации NAD+ и NADPH от возраста в клетках PBMC человека [5]

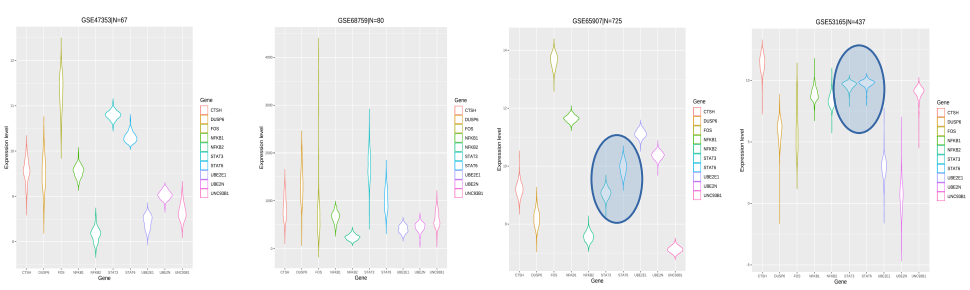
Распределение величин экспрессии

Одни и те же гены в разных датасетах в распределены по-разному, их соотношение различно, поэтому мы реализовывали гармонизацию отдельно для каждого гена

Данные для мужчин



Данные для женщин



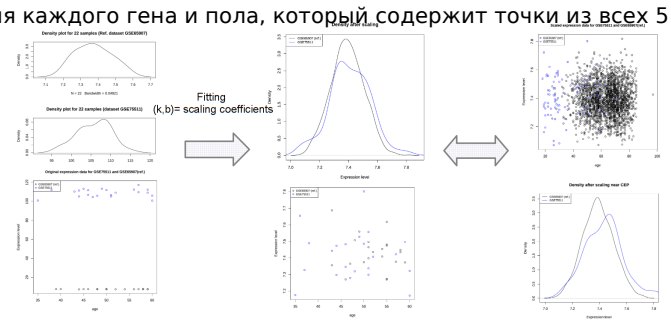
Процедура гармонизации

- 4 датасета (i=1:4) + 1 отдельный («опорный») датасет. Для гармонизации точки для мужчин и женщин объединены.
- Вычисляем 4 “возраста пересечения датасетов” (CEP_i) с помощью пересечения графиков распределения от возраста для каждого датасета, i=1:4). Это нужно, чтобы датасеты выравнивались только по тем точкам, которые расположены в «зоне пересечения» этих датасетов по возрасту.
- Выбираем 28 точек ближайших к CPE_i (слева и справа от CPE_i) для i-го и опорного датасета GSE65907 → получаем два множе ства по 28 точек: EL_{ref}, EL_i
- Сэмплируем 22 из 28 точек (20 раз в цикле) и ищем пары коэффициентов линейного шкалирования (k_i,b_i) согласно алгоритму:

Δ_{1i}=quantile(0.05,EL_{ref})- (k_i*quantile(0.05,EL_i)+b_i)
Δ_{2i}=quantile(0.95,EL_{ref})- (k_i*quantile(0.95,EL_i)+b_i)
rss = sqrt(Δ_{1i}^2+Δ_{1i}^2)
rss→min

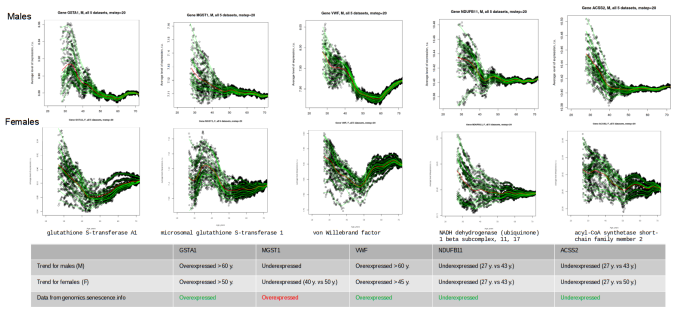
Где quantile(0.05,...), quantile(0.95,...) - соответствующие квантили указанных распределений. При минимизации используется функция *optim* (язык R), старт оптимизации - со случайной пары параметров. Таким образом, получаем 20 пар (k_i,b_i), i=1:4

- Используя эти пары значений (k_i,b_i) получаем 20 вариантов объединённых датасетов для каждого гена и пола, который содержит точки из всех 5 датасетов



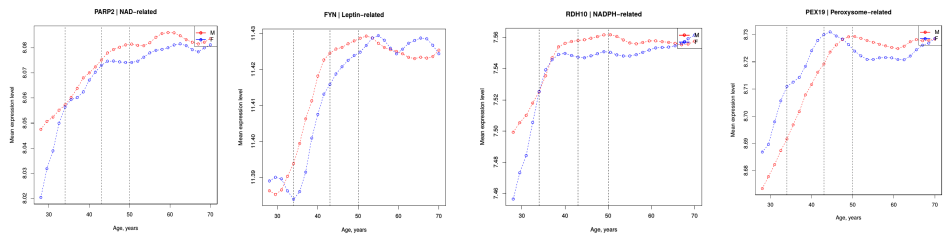
Получение трендов и валидация

Используя модифицированную функцию скользящего среднего (ширина окна усреднения фиксирована «в годах») мы получаем предварительные тренды (чётные точки на графике), которые затем сглаживаются с помощью кубического сплайна (зелёные кривые). Параметр “**mstep**“ определяет ширину окна усреднения. Окончательный тренд определяется с помощью усреднения полученных 20 вариантов сплайнов по «предварительным» трендам (красная кривая).

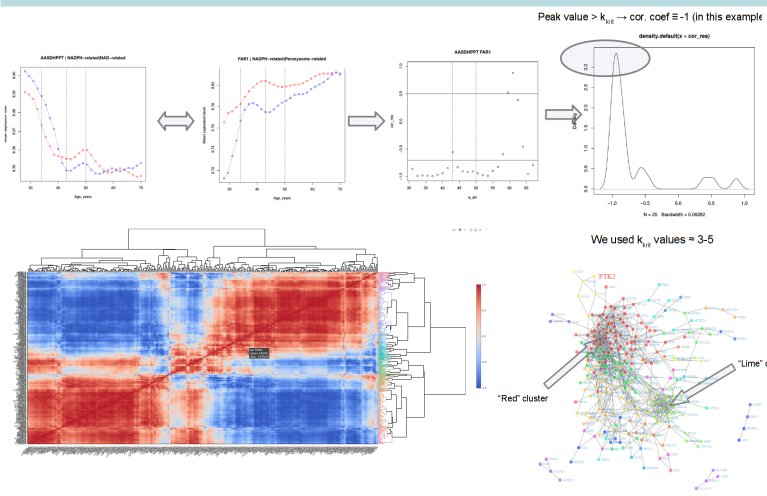


	GSTA1	MGST1	VWF	NDUFB11	ACSS2
Trend for males (M)	Overexpressed >60 y.	Underexpressed	Overexpressed >60 y.	Underexpressed (27 y. vs 43 y.)	Underexpressed (27 y. vs 43 y.)
Trend for females (F)	Overexpressed >50 y.	Underexpressed (40 y. vs 50 y.)	Overexpressed >45 y.	Underexpressed (27 y. vs 43 y.)	Underexpressed (27 y. vs 50 y.)
Data from genomics senescence.info	Overexpressed	Overexpressed	Overexpressed	Underexpressed	Underexpressed

Для валидации мы сравнили поведение некоторых генов с базой GeneAge. Также, для многих генов наблюдается синхронное поведение трендов для мужчин и женщин Поскольку эти наборы точек разные и обрабатывались независимо, их сопоставление также может считаться валидацией нашего подхода.



Кластеризация трендов



Предложен алгоритм детектирования сильно скоррелированных трендов.

В графе связей ребра означают сильную корреляцию трендов экспрессии генов (положительная и отрицательная корреляции).

Мы провели поиск «сообществ» на графах, полученных на основе данных мужчин и женщин с использованием функций библиотеки iGraph R язык.
Для построения интерактивных графиков использовались функции из пакета d3Network R.

Также мы использовали иерархическую кластеризацию для «визуального» определения оптимального количества кластеров (5-6 шт.)

Выводы

В работе получено большое количество (11145 шт.) трендов экспрессии генов PBMC человека при старении, разработаны подходы кластеризации трендов. Проанализированы тренды групп метаболически-значимых генов. Были найдены гены, тренды которых имеют немонотонное поведение в области 43 года у мужчин (метаболического переход) и 55 лет (менопауза) у женщин. Мы полагаем, что данная работа имеет важное значение для поиска генов, тренды которых отражают различные аспекты старения, что может дать важные исходные данные для построения системно-биологических моделей старения.

Литература
1. Lehallier, B., Gate, D., Schaum, N. et al. Undulating changes in human plasma proteome profiles across the lifespan. Nat Med 25, 1843–1850 (2019)
2. Global metabolic profiling to model biological processes of aging in twins Aging Cell. 2019;00:e13073.
3. Rhee, Eun-Jung, et al. "2018 Guidelines for the Management of Dyslipidemia in Korea." Journal of Lipid and Atherosclerosis 8.2 (2019): 78-131.
4.Alpert A. et al. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring //Nature medicine. – 2019. – С. 1.
5. Clement J. et al. The plasma NAD+ metabolome is dysregulated in “normal” aging //Rejuvenation research. – 2019. – Т. 22. – №. 2. – С. 121-130