МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

Биологический факультет
Кафедра биоинженерии

АНАЛИЗ ФИЛЕТИЧЕСКОГО РАСПРЕДЕЛЕНИЯ ГИСТОНОВЫХ ВАРИАНТОВ В ЭУКАРИОТИЧЕСКИХ ОРГАНИЗМАХ

PHYLETIC DISTRIBUTION OF HISTONE VARIANTS IN EUKARYOTIC ORGANISMS

Выпускная квалификационная работа
бакалавра
студента 4 курса А.М. Неугодов

Научный руководитель:
в.н.с. кафедры биоинженерии, к. ф.-м. н.
А.К. Шайтан

Москва
2020

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Гистоны - обширный класс ядерных белков, выполняющих две основные функции: они участвуют в упаковке нитей ДНК в ядре и в эпигенетической регуляции таких ядерных процессов, как транскрипция, репликация и репарация. Гистоновые белки разделяют на репликационно зависимые (канонические) и репликационно независимые, экспрессирующиеся в ходе всего клеточного цикла. Последние также называются гистоновыми вариантами.

Гистоновые варианты крайне разнообразны по своим свойствам и происхождению. Они играют чрезвычайно важную роль в жизнедеятельности эукариотических организмов, а наличие структурированных данных об их разнообразии необходимо для полноценного понимания широкого спектра явлений, изучаемых многими специалистами в области наук о жизни. Для большого количества таксонов разных уровней выявлены уникальные семейства гистоновых вариантов.

В настоящее время стремительно развиваются методы секвенирования. Вследствие этого растет и количество известных белковых последовательностей. Одновременно с этим растет и интерес к механизмам эпигенетической регуляции, центральную роль в которых играют гистоны. Их последовательности, полученные от представителей различных групп организмов, чрезвычайно разнообразны. Это приводит к различиям в особенностях эпигенетической регуляции у разных таксонов. В связи с этим обширный анализ распределения гистоновых вариантов по таксонам является весьма актуальной задачей.

ЦЕЛИ ЗАДАЧИ

Целью данной работы является выявление особенностей распределения канонических гистонов и гистоновых вариантов по избранным таксонам эукариот на основе анализа предварительной версии базы данных гистоновых последовательностей HistoneDB 3.0.

Для достижения данной цели были поставлены следующие задачи:

1. На основе анализа последовательностей гистонов в базе данных GenBank изучить распределение многообразия гистонов в различных таксонах эукариот.
2. Получить распределения по количеству последовательностей всех гистоновых вариантов для четырех наиболее представленных таксонов старшего уровня.
3. Проверить расходящиеся с ожидаемыми результаты анализа филетического распределения с помощью детального сравнения последовательностей методом множественных выравниваний и построения филогенетических деревьев.
4. Изучить распределение последовательностей гистонов по длинам.
5. Изучить распределение по длинам последовательностей для четырех наиболее представленных таксонов старшего уровня.

ОБЗОР ЛИТЕРАТУРЫ

1. Современные представления о макросистематике эукариот

Эукариоты (от др.-греч. εὖ - «полностью», κάρυον - «орех; ядро») – один из трех монофилетических макродоменов живых организмов в трехдоменной системе (два других – бактерии и археи – относятся к прокариотам и не имеют оформленного ядра и мембранных органелл) (Woese et al., 1990). Главным отличительным признаком его представителей является наличие оформленного ядра и мембранных органелл, таких как митохондрии и аппарат Гольджи, и, кроме того, некоторые клетки растений и водорослей содержат хлоропласты. В отличие от обычно одноклеточных архей и бактерий, эукариоты также могут быть многоклеточными и включать организмы, состоящие из множества типов клеток, образующих различные виды тканей. Эукариоты могут размножаться как бесполым путем с помощью митоза, так и половым путем через мейоз и слияние гамет. Другой отличительной особенностью эукариот является механизм компактизации ДНК, основанный на её взаимодействии с гистонами (у Archaea гистоны устроены отлично от эукариотических (Reeve, 1997)).

Основы современной классификации эукариот были заложены в 2005 году Международным Сообществом Протистологов (International Society of Protistologists) (Adl et al., 2005). Домен был разделен на шесть, как предполагалось, монофилетических супергрупп. Однако в 2012 году классификация была пересмотрена, современным сообществом признается пять супергрупп: Amoebazoa, Opisthokonta, Excavata, SAR (Stramenopiles, Alveolata, Rhizaria) и Archaeplastida (Adl et al., 2012). Также существует ряд таксонов, не относящихся к этим супергруппам, или тех, чье положение не определено. Общая схема принятой системы представлена в таблице 1.

Филогенетические деревья эукариот постоянно пересматриваются. В связи с этим мы остановимся только на одном исследовании, которое является достаточно показательным и отражает общие тенденции. На рисунке 1 представлена версия филогенетического дерева эукариот на основе исследований (Burki, 2014).

Таблица 1. Классификация эукариот на уровне старших таксонов (Adl et al., 2012).

		Супергруппы		Представители
Eukaryota	Amorphea	Amoebozoa		Tubulinea Mycetozoa
		Opisthokonta		Fungi Choanomonada Metazoa
				Apusomonada
				Breviata
		Excavata		Metamonada Malawimonas Discoba
	Diaphoretickes			Cryptophyceae
				Centrohelida
				Telonemia
				Haptophyta
		SAR	Rhizaria	Cercozoa Foraminifera Radiolaria
				Alveolata Stramenopiles
		Archaeplastida		Glaucophyta Rhodophyceae
			Plantae	Rhodophyta Viridiplantae
	Incertae sedis Eukaryota			Incertae sedis

В существующей системе ранги (четко обособленные уровни в систематическом положении, имеющие устоявшиеся наименования) частично потеряли прежний смысл, однако для удобства все еще используются. Относительное указание таксонов для группы содержательнее абсолютного, так как отражает объективные представления о соподчиненности групп.

Перечисленным супергруппам обычно не присваиваются именные ранги. Группам, отмеченным в таблице 1 как «представители» и находящимся на ветвях с близким порядком ветвления обычно присваиваются ранги от типа (Radiolaria, Foraminifera) (Margulis, Schwartz,1998) до царства (Fungi, Metazoa) (Simpson et al., 2004). Таким образом, можно видеть, что упоминание рангов без соотнесения со старшими и младшими рангами отражает лишь субъективные представления

об уровнях систематической организации. Поэтому в дальнейшем мы будем использовать ранговую систему только в относительном виде.

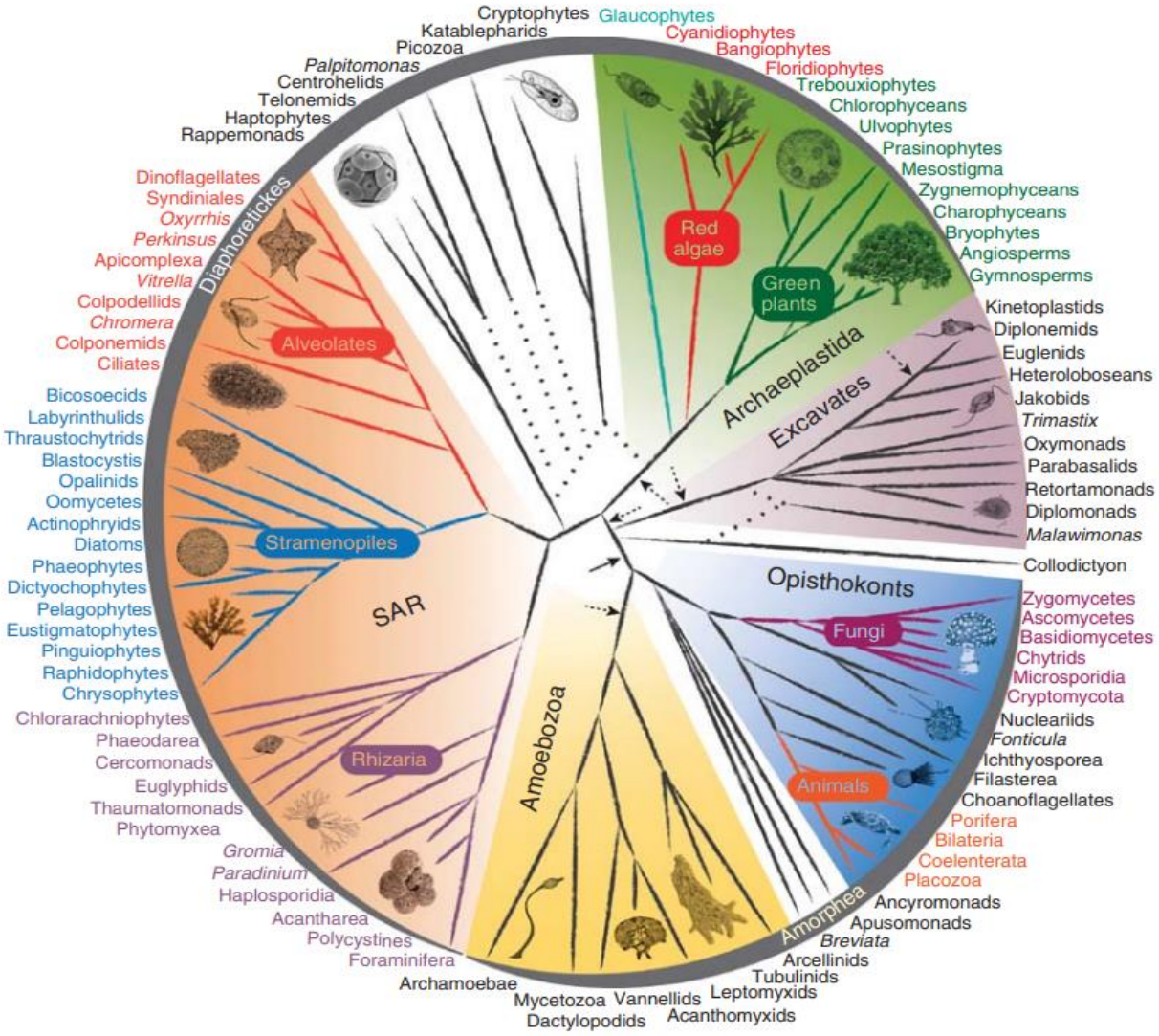

Рисунок 1. Глобальное древо эукариот на основе обширных филогенетических данных (Burki, 2014).

2. Гистоны

2.1. Гистоны – основные белки хроматина

Как уже известно, наследственная информация любой из живых клеток закодирована в гигантских полимерных молекулах ДНК, содержащихся в цитоплазме у доядерных организмов и в ядре, митохондриях и хлоропластах у эукариот. Однако, длина цепи ДНК только хромосомы 1 человека составляет около 8,5 см, в то время как диаметр клеточного ядра составляет примерно 10

мкм (Watson, Crick, 1953). Данный факт подразумевает необходимость существования способа упаковки, компактизации и упорядочения клеточной ДНК. Для этой цели в клетке существуют особые белки, гистоны, способные связывать и конденсировать ДНК до размеров, сопоставимых с диаметром клеточного ядра.

Гистоны – это небольшие, сильно основные, положительно заряженные благодаря высокому содержанию лизина и аргинина, белки, составляющие основную часть белковой фракции хроматина. Гистоновые белки являются относительно небольшими белками длиной 102-135аа и молекулярной массой примерно 10-15 kDa, которые имеют общий структурный мотив, называемый «histone fold» (DeLange, Smith, 1971). Стоит отметить, что так называемый линкерный гистон H1 отличается по размерам, свойствам и происхождению от остальных гистоновых белков: его длина превышает 200 аминокислотных остатков, а молекулярная масса составляет примерно 20 kDa (DeLange, Smith, 1971). Однако существуют различные варианты гистоновых белков,

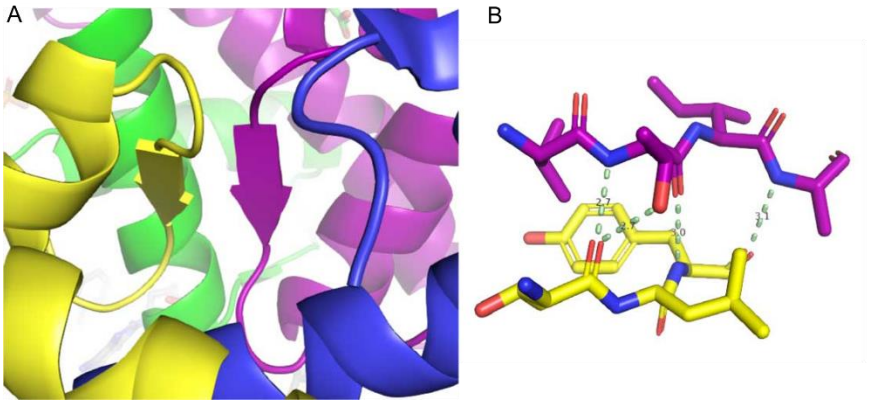

Рис. 2. A – β-мост, связывающий H2A (фиолетовым) и H4 (желтым) гистоны. B – водородные связи β-моста и их длины в ангстремах. PDB ID: 4QLC (Zhou et al., 2015).

характеристики которых могут несколько варьироваться. Одним из таких гистонов является гистоновый вариант macroH2A длиной, в среднем, 368 аминокислотных остатков и молекулярной массой около 42 kDa, обладающий N-концевым H2A-подобным доменом и C-концевым глобулярным доменом (Pehrson et al., 1992, Chakravarthy et al., 2005).

Histone-fold мотив образован тремя α-спиралями, соединенными двумя петлями, которые содержат β-структуры (рис. 2A) и обеспечивают межмолекулярные взаимодействия между белками – β-мосты (рис. 2B) (Arents et al., 1993). Стоит отметить, что положительно заряженные аминокислоты сосредоточены главным образом на C- и N-концах. Особенно важно то, что N-концевые участки неструктурированы и выступают за пределы октамера, обеспечивая связывание ДНК и возможность регуляции особенностей структуры хроматина и экспрессии генов посредством посттрансляционных модификаций гистонов.

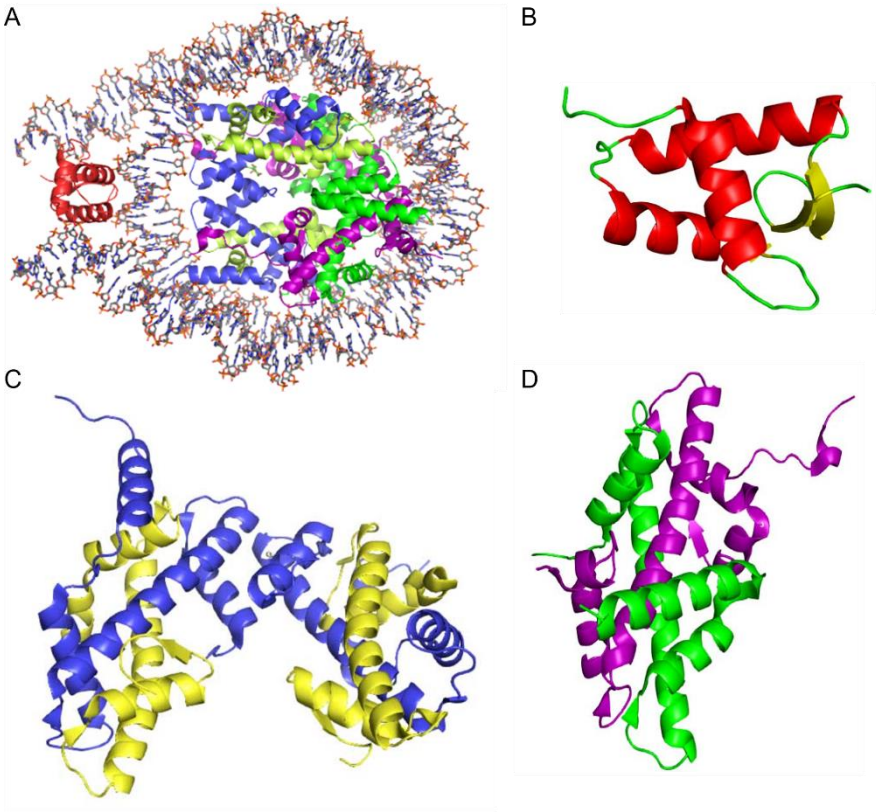

Рис. 3. A – Структура хроматосомы в разрешении 3,5 Å. Фиолетовым – H2A, (*Drosophila melanogaster*) зеленым – H2B (*D. melanogaster*), синим – H3 (*D. melanogaster*), желтым – H4 (*D. melanogaster*), красным – птичий H1.0 (*Gallus gallus*). B – трехмерная структура гистонового белка H1.0 птичьих эритроцитов (Talbert et al., 2012). C – трехмерная структура тетрамера $(H3/H4)_2$. D – трехмерная структура димера H2A/H2B. PDB ID: 4QLC (Zhou et al., 2015).

На сегодняшний день выделяют пять основных типов гистонов: так называемые линкерный H1 и коровые (от англ. «core» - ядро) H2A, H2B, H3 и H4, образующие стержень нуклеосомы – гистоновый октамер (рис. 3A), состоящий из двух H2A/H2B димеров (рис. 3D) и тетрамера $(H3/H4)_2$ (рис. 3C)

через, так называемое, «handshake-interaction» (Oudet et al., 1975, Thomas, Kornberg, 1975, Eickbush, Moudrianakis, 1978, Arents et al., 1991). Цепь ДНК, в свою очередь, взаимодействуя с повторяющимися положительно заряженными ДНК-связывающими структурными мотивами, обвивается вокруг гистонового октамера в 1,8 витка, причем длина этого сегмента ДНК составляет примерно 146 п.н. (пар оснований) (Oudet et al., 1975, Arents et al., 1993). В свою очередь линкерный гистон H1, не являясь составной частью гистонового октамера, взаимодействует с нуклеосомой и 20 п.н. линкерной ДНК и участвует в последующей суперспирализации ДНК и образовании зигзагообразных или соленоидных хроматиновых 30 нм-фибрилл (Tremethick, 2007, Öztürk et al., 2020). Таким образом комплекс нуклеосомы с гистоном H1, хроматосома (рис 3А), обвивает вокруг себя 166 п.н. ДНК в 2 витка (Simpson, 1978).

Как уже указывалось выше, важную роль в структурной организации играют также посттрансляционные модификации N-концевых участков гистонов, выступающих за пределы нуклеосом. Изменения в структуре хроматина вследствие посттрансляционных модификаций гистонов происходят из-за изменения заряда гистонов, нарушения внутринуклеосомных взаимодействий, а также из-за рекрутирования регуляторных белков (Bannister, Kouzarides, 2011). К основным видам модификаций можно отнести ацетилирование (по ε-аминогруппе остатков лизина), фосфорилирование (по серину, треонину и тирозину), метилирование (по аргинину и ε-аминогруппе остатков лизина), убиквитинирование (по лизину) и др. (Bannister, Kouzarides, 2011, Zhao et al., 2019). Ацетилирование и фосфорилирование гистонов снижает общий заряд октамерного стержня, что в свою очередь приводит к дестабилизации нуклеосомного комплекса и декомпактизации хроматина. Это, например, облегчает доступ транскрипционного аппарата к ДНК (Kiefer et al., 2008, Bannister, Kouzarides, 2011, Zhao et al., 2019). Однако, существуют и исключения, как, например, повсеместное фосфорилирование H3S10 во время митотического деления ассоциировано с усилением конденсации хроматина, что может быть связано с необходимостью диссоциации гетерохроматинового белка

HP-1 с ДНК (Fischle et al., 2005, Bannister, Kouzarides, 2011). В то же самое время рекрутирование HP-1 к ДНК обеспечивается триметилированием H3K9 в теломерных и перицентральных регионах хромосом, а также в зонах репрессии генов (Fischle et al., 2005). Важно также отметить, что для согласования исследований в области биологии гистонов и их посттрансляционных модификаций была составлена номенклатура Брно (Brno nomenclature), которая повсеместно используется и на сегодняшний день (Turner, 2005). Данная номенклатура обеспечивает унифицированное письменное отображение шести типов модификаций: ацетилирование, метилирование, фосфорилирования, убиквитилирование, сумоилирование и АДФ-рибозилирование (Turner, 2005).

2.2. Гистоновые варианты

Как уже отмечалось выше, помимо «канонических» форм гистоновых белков существуют также «неканонические» гистоновые варианты – гистоны, кодируемые неаллельными генами, встраивающиеся в ДНК по репликационно-независимому механизму и конститутивно экспрессируемые во время всего клеточного цикла (возможно и в процессе митоза) (Palozola et al., 2017, Singh et al., 2018). В отличие от генов канонических гистонов, располагающихся в определенных локусах хромосом и сохраняющих свою консервативность на протяжение эволюции, гены неканонических гистонов «разбросаны» по геному и могут сильно варьироваться в зависимости от видовой принадлежности организма (Singh et al., 2018). Эти особенности обусловлены тем, что упаковка ДНК не является основной функцией этих гистонов, в отличие от «канонических» гистоновых белков. Было показано, что гистоновые варианты могут участвовать в огромном множестве физиологических и патологических процессов, среди которых оплодотворение, инактивация Х-хромосомы, сегрегация хромосом, регуляция транскрипции, образование гетерохроматина, репарация ДНК и др. (Skene, Henikoff, 2013). Далее будут рассмотрены некоторые наиболее изученные на данный момент гистоновые варианты, а также их роль в различных клеточных процессах.

Каноничные гистоны кодируются 72 генами, собранными в несколько кластеров тандемно повторяющихся копий генов, наибольший из которых HIST1 у человека располагается на хромосоме 6 и состоит из 55 генов, два кластера на хромосоме 1 HIST2 (12 генов) и HIST3 (4 гена) и один кластер HIST4 (1 ген) на хромосоме 12. Интересно, что для обеспечения жесткой регуляции транскрипции генов канонических гистонов их пре-мРНК не подвергаются 3'-полиаденилированию, а формируют 3'-шпильки, которые также эволюционно консервативны и позволяют осуществлять регуляцию их метаболизма посредством особого белка SLBP (stem loop binding protein) (Marzluff et al., 2008). Тем не менее, из этого правила существуют и исключения: было показано, что в терминально дифференцированных клетках пре-мРНК 10 генов кластеров HIST1 и HIST2 проходили сплайсинг и 3'-полиаденилирование (Lyons et al., 2016).

В отличие от канонических гистонов, которые экспрессируются главным образом в S-фазу клеточного цикла и встраиваются по репликационно-зависимому механизму, гистоновые варианты экпрессируются конститутивно и встраиваются в хроматин по репликационно-независимому механизму в процессе не только удвоения ДНК, но и в процессе транскрипции, изменяя структуру нуклеосомы и организацию хроматина. Также их экспрессия регулируются независимо от механизмов регуляции экспрессии канонических гистонов, так как мРНК гистоновых вариантов имеет 3'-поли-(А) последовательность (Marzluff et al., 2008). Встраивание гистонов в нуклеосомные комплексы осуществляется посредством белков-шаперонов. Например, шаперон HIRA обеспечивает сборку нуклеосом в процессе репликации ДНК, перенося на ДНК тетрамеры, содержащие гистоновый вариант H3.3 (Tagami et al., 2004).

Несмотря на высокую структурную схожесть с каноническими H3, H3.3 выполняет некоторые функции, выходящие за рамки компактизации ДНК. В ходе многих исследований было показано, что гистон H3.3 встраивается в ДНК в промоторных и энхансерных областях, в регионах терминации транскрипции, а также гетерохроматиновых областях (Skene, Henikoff, 2013, Talbert, Henikoff,

2017). Его нокдаун приводит ко многочисленным нарушениям, в том числе к транспозиции эндогенных ретровирусов и их дерепрессии, нарушению структуры и сестринскому хроматидному обмену теломерных областей (Udugama et al., 2015, Elsässer et al., 2015), а нарушение его функции вследствие мутаций приводит к нарушениям сперматогенеза и дифференцировки (Skene, Henikoff, 2013, Talbert, Henikoff, 2017, Gehre et al., 2020).

Особым гистоновым вариантом H3 является гистон cenH3 (CENP-A), специфичный для центромер у животных, а также принимающий участие в образовании кинетохор хромосом (Talbert, Henikoff, 2020). Стоит отметить, что cenH3 имеет достаточно мало сайтов посттрансляционной модификации, сходство аминокислотных последовательностей cenH3 и H3 относительно низко – около 50%, однако структурно cenH3 очень похож на каноничный H3 за исключением L1 петли, которая имеет вставки R80 и G81, необходимые для формирования кинетохорного комплекса (Sharma et al., 2019). Многие исследования показали, что cenH3 эктопически гиперэкспрессируется в клетках многих типов раковых опухолей, в том числе в клетках рака груди и колоректального рака, что придает клеткам устойчивость к воздействию ионизирующей радиации (Sharma et al., 2019).

Об одном из вариантов гистона H2A, макрогистоне macroH2A, выше уже было упомянуто. Его гистон-подобный домен обладает высоким структурным сходством с каноническим H2A, однако он обладает также глобулярным доменом, выступающим из структуры нуклеосомы и обладающим структурой, отличной от histone-fold (Pehrson et al., 1992, Chakravarthy et al., 2005). Также стоит отметить, что экспрессируется главным образом в соматических клетках (Chadwick, Willard, 2001, Barrero et al., 2013). Многие исследования показали связь этого гистонного варианта с генетической репрессией, в том числе была показана ко-локализация macroH2A с маркерами гетерохроматина инактивированной X-хромосомы (Xi) РНК Xist и H3K27me3 (Chadwick, Willard, 2004). Как выяснилось в ходе исследований, нокдаун macroH2A приводит к нарушению дифференцировки человеческих ESCs (embryonic stem cells), а также

является серьезным препятствием для создания индуцированных плюрипотентных клеток путем транскрипционного репрограммирования (transcription-based reprogramming) клеток (Barrero et al., 2013, Barrero et al., 2013a).

Другой вариант H2A, гистон H2A.Z, также может встраиваться в нуклеосомный комплекс независимо от репликации (Skene, Henikoff, 2013). Интересно, что последовательность H2A.Z сходна с последовательностью каноничного H2A лишь на 60%, что дает повод предположить наличие уникальной функции у данного варианта и характерных отличиях в его структуре (Suto et al., 2000). К таким характерным отличиям можно отнести возможность связывания иона металла (марганца, меди или цинка), расширенный кислотный лоскут и ослабление взаимодействия с $(H3/H4)_2$ тетрамером (Suto et al., 2000). Такие особенности обуславливают участие во множестве различных процессов, в том числе отмечено, что H2A.Z участвует как в активации, так и в репрессии транскрипции благодаря тому, что фланкирует область начала транскрипции, также он обнаруживается и в энхансерных регионах, а его нокдаун приводит к нарушению процесса митоза. Также данный гистоновый вариант располагается в гетерохроматиновых теломерных и перицентральных областях хромосом, предотвращая распространение гетерохроматина (Skene, Henikoff, 2013, Talbert, Henikoff, 2017).

Гистоновый вариант H2A.X в сравнении с каноничным гистоном H2A обладает важной чертой – наличием SQ(E/D)(I/L/Y) последовательности. Эта последовательность необходима для обеспечения жизненно важного процесса для клетки – репарации двуцепочечных разрывов (Pinto, Flaus, 2010, Skene, Henikoff, 2013). Двуцепочечный разрыв приводит к фосфорилированию H2A.X киназами ATM, ATR, DNA-PK по S139 в радиусе около 2Mb от места разрыва, что позволяет локализовать место разрыва и амплифицировать сигнал комплексам репарации ДНК (Pinto, Flaus, 2010, Skene, Henikoff, 2013).

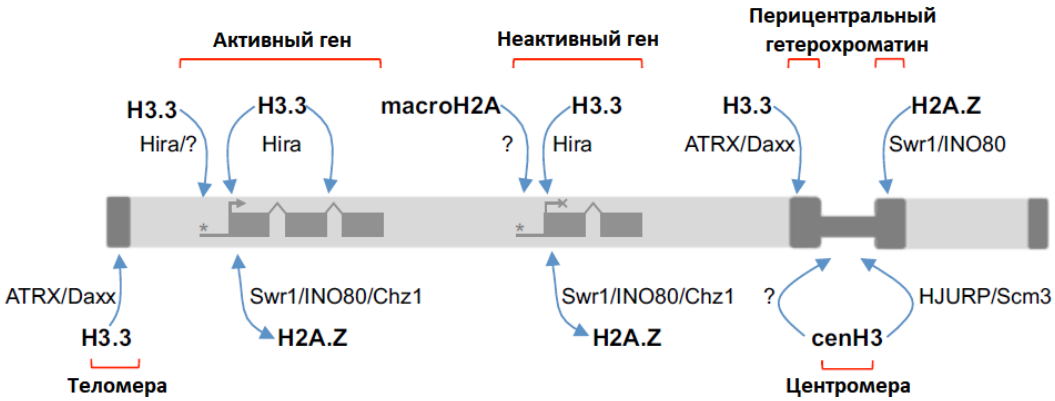

Рис. 4. Схема геномной локализации гистоновых вариантов, для каждого из которых указан шаперон, участвующий в его встраивании. (Skene, Henikoff, 2013).

2.3. Существующая номенклатура гистоновых вариантов

Поскольку число гистоновых вариантов достаточно велико, а с удешевлением и рутинизацией геномных исследований число секвенируемых многократно возросло, встала острая необходимость в создании унифицированной системы номенклатуры вариантов гистоновых белков для использования в исследованиях, а также в базах данных, как например в HistoneDB 2.0. На данный момент используется номенклатура гистоновых вариантов, предложенная Talbert и коллегами в 2012 году (Talbert et al., 2012, Draizen et al., 2016).

Основной причиной необходимости внедрения данной номенклатуры авторы сочли неструктурированность и беспорядочность наименований гистоновых белков. С самого начала числовые суффиксы (например, H3.1) присваивались произвольно после их хроматографического разделения, либо в зависимости от белкового семейства и электрофоретической подвижности в различных условиях, либо в порядке открытия. Также применялись и буквенные суффиксы с или без пунктуационных знаков для минорных представителей семейств белков. Особенная неопределенность возникала с использованием суффикса V для обозначения вариантных форм: гистон H2Bv Plasmodium и гистон H2BV Trypanosoma могут быть очень легко спутаны. Также авторы

отметили отсутствие различий в использовании буквенных и числовых суффиксов, так как и буквы, и цифры использовались для обозначения одних и тех же понятий, как, например, обозначения паралогов или обозначения продуктов альтернативного сплайсинга. Также в номенклатуре использовались префиксы, обозначающие последовательности, подвергшиеся дивергенции. Использовались и видовые префиксы для указания видоспецифичности, зачастую в то же самое время использовались и суффиксы. Не были редкостью и дескрипторы в виде отдельных слов или словосочетаний (Talbert et al., 2012).

Talbert и коллеги в 2012 году предложили новую систему номенклатуры, основанную на филогенетических отношениях среди гистоновых белков. Согласно ей, буквенные префиксы могут использоваться для обозначения структурно различающихся вариантов внутри клады более высокого уровня и должны указываться строчными буквами, в то время как суффиксы прописными, во избежание путаницы при указании посттрансляционных модификаций по номенклатуре Брно. Предложено отказаться от использования некоторых префиксов. Например, вместо наименования CENP-A предлагается использовать наименование с префиксом cenH3 (Talbert, Henikoff, 2013). В общем случае Talbert и коллеги рекомендуют использовать однобуквенные суффиксы. Например, гистоны H2A.Bbd млекопитающих предложено именовать H2A.B. Филогенетический анализ показал, что зависимые от репликации H3.2 и H3.1 и репликационно-независимый H3.3 являются полифилетической группой, в то время как H3 Ascomycetes - H3.3-подобный белок. На основании этого Talbert и коллеги рекомендовали использовать H3 без дескрипторов и суффиксов только в отношении Ascomycetes, а для млекопитающих продолжить использование H3.1, H3.2, H3.3. Использование числовых суффиксов может быть оправдано также для обозначения белков-паралогов, которые должны соответствовать суффиксам белков-ортологов родственных видов. Также предложено отказаться от использования суффикса ‘v’, решая проблему вышеназванного примера названий H2Bv Plasmodium и H2BV Trypanosoma заменой на H2B.Z и H2B.V соответственно, которые находятся в одной кладе и являются родственными

друг другу, но долгое время эволюционировали независимо. Дескрипторы в свою очередь рекомендуется использовать более гибко для указания специфичных вариантов для определенных стадий или видовой специфичности.

Номенклатура также подразумевает систематизацию использования пунктуационных знаков. В случае суффиксов, использование точки для отделения определяющего символа (цифры или буквы) обозначает точку ветвления филогенетического дерева. То же правило применимо и для обозначения вариантов альтернативного сплайсинга (Talbert et al., 2012).

Использование разных наименований для обозначения одних и тех же белков или белков-ортологов не рекомендуется, например, для обозначения Htz1p Saccharomyces, hv1 Tetrahymena и H2Av/H2AvD/D2 Drosophila должно быть использовано наименование H2A.Z (Talbert et al., 2012).

Особое внимание было уделено номенклатуре вариантов гистонов H1, которые вероятно, эволюционировали отдельно и, скорее всего, от бактериальных белков (Kasinsky et al., 2001, Talbert, Meers, Henikoff, 2019). Изначально номенклатура, используемая для обозначения этих белков, использовала строчные буквы, обозначающие порядок выделения из хроматографической колонки Bio-Rex 70 и электрофоретическую подвижность. Для обозначения человеческих гистоновых вариантов H1 использовались суффиксы, различающие H1 соматических (H1.1 H1.2, …, H1.5, H1x, а также H1° или H1.0) и половых клеток (H1t, H1T2, Hils1 – специфичные для яичек, H1oo – специфичные для ооцитов). Однако, эта номенклатура также вводит в заблуждение вследствие несоответствия наименований ортологов среди разных видов. Поэтому рекомендуется именовать такие белки одинаково. Например, поскольку гистон куриных эритроцитов H5 является ортологом H1.0 млекопитающих, рекомендуется использование наименования H1.0 и для гистона H5. Также Talbert и коллеги рекомендовали заменить суффиксы, используемые для обозначения H1 половых клеток, H1t, например, предложено заменить на TS (testis-specific) H1.6, а H1oo заменить на OO H1.8 (Talbert et al., 2012).

2.4. Базы данных гистоновых вариантов

В 1995 году исследователями из National Center for Biotechnology Information (NCBI) была создана Histone Sequence Database, содержавшая более 1300 аминокислотных и нуклеотидных последовательностей гистоновых белков, а в качестве основы для сравнения использовались последовательности человека и курицы (Baxevanis, Landsman, 1996). Она представляла собой объединение данных SwissProt 31.0, PIR 45.0, GenPept 91.0, PDB с использованием алгоритма BLASTP. Важной особенностью было то, что при выравнивании последовательностей гистоновых белков, аминокислоты histone-fold выделялись в отдельную рамку (Baxevanis, Landsman, 1996). Главным нововведением Histone Sequence Database в 1997 году было использование двух различных алгоритмов BLASTP и PSI-BLAST. Аминокислотные последовательности выводились в формате FASTA. Для попарного и множественного сравнения последовательностей использовалась программа CLUSTAL W. Также была создана поисковая машина базы данных. Поиск histone-fold мотивов осуществлялся с использованием Motif Search Tool и PROBE. (Baxevanis, Landsman, 1998). В 1998 году исследователи добавили в базу данных сведения о посттрансляционных модификациях, локусах генов и данные геномных исследований. База данных была разделена на два сета, первый из которых содержал информацию «избыточную» (redundant), а второй содержал информацию о гистоновых вариантах, последовательность которых была определена лишь один раз (non-redundant) (Makalowska et al., 1999).

В 1999 году база данных стала носить название The Histone Database (HDB). Основным нововведением стало разделение на 10 областей: источники и обобщенная информация; поисковая машина; все последовательности эукариот в формате FASTA; набор «избыточных» последовательностей; последовательности гистонов архей и негистоновых белков в формате FASTA; множественные выравнивания коровых, линкерных гистонов, histone-fold регионов гистонов архей и негистоновых белков; таблица 3D-структур гистонов и histone-fold мотивов; информация о посттрансляционных модификациях;

графическое изображение хромосом и локусов генов гистонов; таблица несоответствий в последовательностях (Sullivan et al., 2000).

В 2002 году число видов организмов, последовательности гистонов которых содержались в базе данных HDB, достигло 310, а общее число последовательностей превысило 3000. Также для выравнивания стал применяться CLUSTALX. (Sullivan et al., 2002).

В 2006 году число видов организмов продолжило расти и достигло 975. Множественные выравнивания последовательностей гистонов выполнялись с использованием MUSCLE и CLUSTAL-ALW (Marino-Ramirez et al., 2006).

В 2011 году к источникам был добавлен Protein Research Foundation (PRF), коллекция гистонов была расширена с использованием программы поиска гомологий HMMER3. База данных содержала 7356 таксономических идентификаторов гистонов, что примерно соответствовало количеству организмов (Marino-Ramirez et al., 2011).

В 2015 году произошло масштабное обновление базы данных, которая теперь носит название The Histone Database 2.0. Главным нововведением этой версии стало использование новой филогенетической номенклатуры гистонов, предложенной Talbert и коллегами в 2012 году (Talbert et al., 2012). Также были построены филогенетические деревья для каждого из гистонов и гистоновых вариантов (Draizen et al., 2016).

МАТЕРИАЛЫ И МЕТОДЫ

1. Выбранные для рассмотрения таксоны и принципы их отбора

В данной работе мы анализировали распределение гистоновых вариантов на двух выделенных нами условных уровнях в таксономической системе эукариот. Условные уровни не отражают реальных эволюционных дистанций и порядка ветвления этих групп (см. рисунок 1). Данные таксоны были выбраны, прежде всего, как представляющие все пять супергрупп эукариот, из соображений их популярности и узнаваемости, а также для того, чтобы таксоны младшего уровня соответствовали рангу типа (Metazoa, некоторые группы протистов) или отдела (грибоподобные, растительноподобные организмы). Взаимоотношения выбранных для рассмотрения таксонов представлены в таблице 2.

Так как для обозначения таксономического положения организмов – источников последовательностей применяется система NCBI-taxonomy, то мы в своей работе также опирались на неё. В случае отсутствия в данной системе каких-либо изначально запланированных для рассмотрения групп мы выбирали наиболее полно представляющую группу в составе отсутствующей.

Высший уровень представлен группами Amoebozoa, Metazoa, Fungi, Plantae, Discoba, SAR (Stramenopiles, Alveolata, Rhizaria). Ввиду отсутствия единой позиции в различных источниках о присвоении этим группам определенных рангов в данной работе ранговые наименования для старшего уровня не используются. Некоторые группы являются наиболее известными таксонами в составе супергрупп (Metazoa, Fungi для Opisthokonta, Plantae для Archaeplastida). Группа Discoba бала выбрана как наиболее полно представляющая отсутствующую в системе NCBI-taxonomy супергруппу Excavata.

Для ряда таксонов старшего уровня (Metazoa, Fungi, Plantae, SAR) исследовалось распределение на младшем уровне. Всем группам, составляющим этот уровень, как было сказано выше, обычно присваиваются эквивалентные друг другу ранги типа или отдела.

Рассмотрение таксонов рангом ниже типов и отделов было сочтено нецелесообразным в силу обнаружившейся значительной неточности работы

скрытой марковской модели (см. ниже) при различении гистоновых последовательностей таксонов данных уровней.

Таблица 2. Схема взаимоотношений между таксонами старшего и среднего условных уровней. Знак «-» означает, что для таксонов не проводился анализ распределения гистоновых вариантов или ввиду их малой известности или значимости для человека, или из-за недостатка данных. Пропуски в столбце «Промежуточные описания» означают, что для таксонов условного среднего уровня отсутствуют старшие материнские безранговые группы.

Супергруппа	Условный старший уровень	Промежуточные описания	Условный младший уровень
Unikonta, Amoebozoa	Amoebozoa		-
Unikonta, Opisthokonta	Metazoa	Prometazoa	Тип Porifera
		Eumetazoa	Тип Cnidaria
		Eumetazoa, Protostomia, Ecdysozoa	Тип Arthropoda
			Тип Nematoda
		Eumetazoa, Protostomia, Spiralia	Тип Plathelminthes
			Тип Annelida
			Тип Mollusca
		Eumetazoa, Deuterostomia	Тип Echinodermata
			Тип Chordata
	Fungi	Низшие грибы	Отдел Chytridiomycota
			Отдел Zygomycota
		Подцарство Dikarya	Отдел Ascomycota
			Отдел Basidiomycota
Bikonta, Archaeplastida	Plantae		Отдел Rhodophyta
			Отдел Chlorophyta
			Отдел Charophyta
			Bryophyta
			Tracheophyta
Bikonta, SAR	SAR	Stramenopiles	Отдел Ochrophyta
			Отдел Oomycota
		Alveolata	Тип Ciliophora
			Тип Apicomplexa
			Тип Dinoflagellata
		Rhizaria	-
Bikonta, Excavata	Discoba		-

2. Метод множественных выравниваний

Для детального сравнения последовательностей гистоновых белков мы использовали метод множественного выравнивания последовательностей.

Выравнивание последовательностей (sequence alignment, SA) - метод, призванный идентифицировать области сходства, которые могут отражать функциональные, структурные или эволюционные взаимоотношения между последовательностями. Обычно представляется как расположение последовательностей ДНК, РНК или белка друг над другом с выделением областей сходства, а также пробелов (гэпов) в одной последовательности относительно другой знаком “-”. Множественное выравнивание последовательностей (sequence alignment, MSA) - выравнивание трех или более последовательностей белка, ДНК или РНК. Во многих случаях предполагается, что входной набор последовательностей имеет эволюционную связь. Поскольку три или более последовательности схожей длины почти всегда требуют много времени для выравнивания вручную, для получения и анализа выравниваний используются вычислительные алгоритмы.

В настоящей работе мы используем программу (и соответствующий алгоритм MUSCLE (Edgar, 2004). Она относится к так называемым итерационным методам множественных выравниваний. В данных методах окончательное выравнивание строится путем объединения парных выравниваний от более похожих пар к менее похожим. Этот процесс происходит в два этапа: представление отношений между последовательностями в виде направленного графа – направляющего дерева и построения MSA посредством присоединения последовательностей в соответствии с направляющим деревом. При этом исходные последовательности многократно перераспределяются. Поэтому итерационные методы в общем случае дают более точный результат, чем прогрессивные, в которых такого перераспределения не происходит, и ошибки, совершенные на начальном этапе выравнивания, сильно влияют на конечный результат.

MUSCLE (MUltiple Sequence Comparison by Log-Expectation, множественное сопоставление последовательностей по логарифмическому ожиданию) (Edgar, 2004) - программа, реализующая один из итерационных методов множественных выравниваний. Она основана на подсчете расстояний двумя способами: с помощью сравнения k-меров (для начального быстрого выравнивания) и с использованием модели расстояний Кимуры (для уже выровненных последовательностей). Алгоритм MUSCLE проходит в три этапа: черновой прогрессивный, улучшенный прогрессивный и уточняющие этапы. На рисунке 5 представлен общий алгоритм работы MUSCLE (Edgar, 2004).

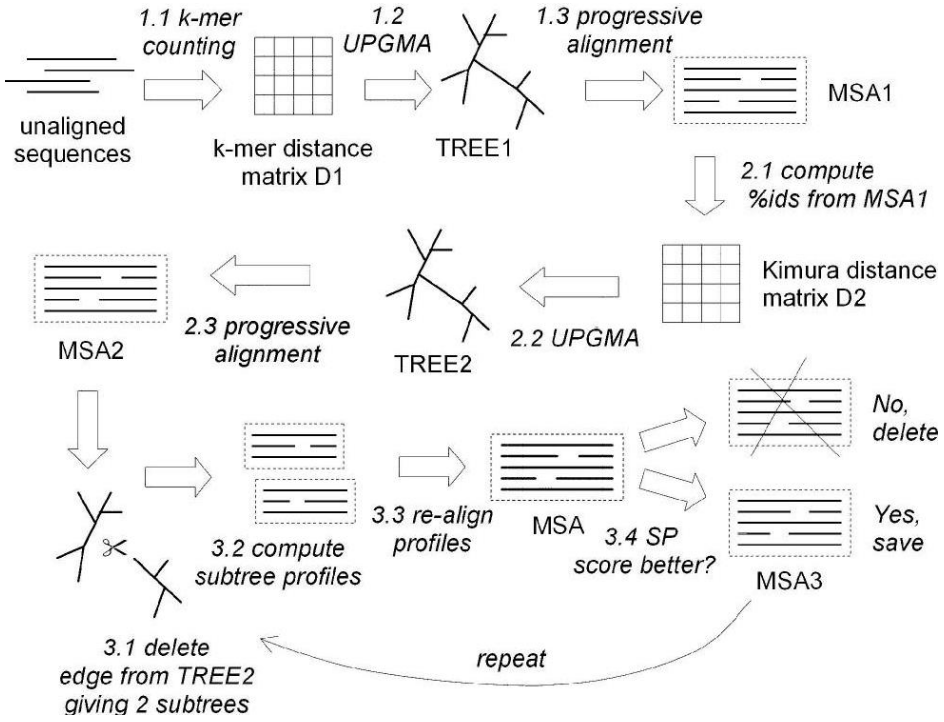

Рис. 5. Схематическое изображение алгоритма работы MUSCLE (Edgar, 2004)

Целью первого этапа является создание чернового множественного выравнивания с преобладанием скорости над точностью. Для каждой пары входных последовательностей вычисляется расстояние на основе подсчёта совпадающих k-меров, что дает матрицу расстояний D1. Матрица D1 кластеризуется UPGMA, создавая двоичное дерево TREE1. Затем строится выравнивание по порядку ветвления TREE1. На каждом листе профиль строится из входной последовательности. Узлы в дереве учитываются в порядке префикса

(дочерние элементы перед их родителями). На каждом внутреннем (не конечном) узле попарное выравнивание строится из двух дочерних профилей, давая новый профиль, который назначается этому узлу. В корне получается итоговый профиль выравнивания MSA1.

Основной причиной ошибок в черновом прогрессивном этапе является неточность меры расстояния, основанной на сравнении k-меров, что приводит к неоптимальному дереву. Поэтому MUSCLE переоценивает дерево, используя расстояние Кимуры, которое является более точным, но требует выравнивания. Расстояние Кимуры (Kimura, 1991) для каждой пары входных последовательностей вычисляется из MSA1, давая матрицу расстояний D2. Матрица D2 кластеризуется UPGMA аналогично тому, как это происходит на первом этапе, создавая двоичное дерево TREE2. Последовательное выравнивание производится по дереву TREE2 (аналогично первому этапу), профиль множественного выравнивания MSA2. Оптимизация происходит посредством построения выравниваний только для поддеревьев, чьи порядки ветвления изменились относительно TREE1.

Затем рассматриваются ребра дерева TREE2 в порядке уменьшения расстояния от корня. TREE2 делится на два поддерева путем удаления ребра. в каждом поддереве строится профиль множественного выравнивания. Новое множественное выравнивание производится путем повторного выравнивания двух профилей. Если оценка SP (pare score) улучшена, новое выравнивание сохраняется, в противном случае оно отбрасывается. Третий этап повторяются до получения оптимального дерева или до достижения определенного пользователем предела. В нашей работе мы использовали режим 10 итераций. Иначе говоря, этапы третьей части алгоритма повторялись десятикратно.

3. Используемая база данных

В качестве источника данных используется таблица в виде .csv файла, общий вид которой представлен в таблице № 3.

Таблица 3. Общий вид таблицы, созданной на основе БД HistoneDB 2.0. В столбце «id» представлены идентификаторы последовательностей, в столбце «taxonomy_id» - идентификаторы систематического положения объекта в NCBI Taxonomy, в столбце «variant_id» - идентификаторы наименования варианта или канонического гистона; «header» - краткое описание последовательностей; «sequence» - последовательность аминокислотных остатков в однобуквенном коде.

id	taxonomy_id	variant_id	header	sequence	canonical
1AOI_B	8355	canonical_H4	1AOI_B Chain B,…	KRHRKVL…	false
1AOI_C	8355	canonical_H2A	1AOI_C Chain C,…	GKQGGKT…	false
1AOI_D	8355	canonical_H2B	1AOI_D Chain D,…	KKRRKTR…	false
…	…	…	…	…	false

Таблица была сформирована на основе рабочей версии проекта по обновлению базы данных гистоновых вариантов HistoneDB 2.0 - with Variants (Draizen et al., 2016). Данный проект осуществляют сотрудники кафедры биоинженерии Биологического факультета МГУ имени М.В. Ломоносова: в.н.с., к.ф.-м.н. А.К. Шайтан и аспирант Л. Сингх. Он является логическим продолжением и развитием работы 2016 года по созданию названной базы данных с помощью скрытых марковских моделей (Draizen et al., 2016). Отдельно стоит уточнить, что автор текущей работы лишь анализировал уже созданную сотрудниками кафедры базу данных и проверял точность её работы, но не участвовал в её создании.

Представленные в таблице последовательности по способу отбора разделены на две группы. Небольшая их часть называется «курируемым набором», т.е. отобранным вручную последовательностями гистоновых белков. В ходе проекта по обновлению сотрудники кафедры не меняли последовательности в составе курируемого набора и общие принципы работы модели, поэтому далее речь будет вестись на основании статьи (Draizen et al., 2016).

Основная часть материала представляет собой так называемый автоматически сгенерированный набор, сформированный на основе

курируемого с использованием скрытых Марковских моделей (Hidden Markov models, HMM). Этот набор был извлечен из не избыточной («non-redundant», nr) базы данных белковых последовательностей, поддерживаемой NCBI (National Center for Biotechnological Information, USA). Термин «не избыточный» в данном случае означает, что идентичные последовательности сгруппированы в виде единичных записей. Общая схема работы HistoneDB 2.0 - with Variants представлена на рисунке 6.

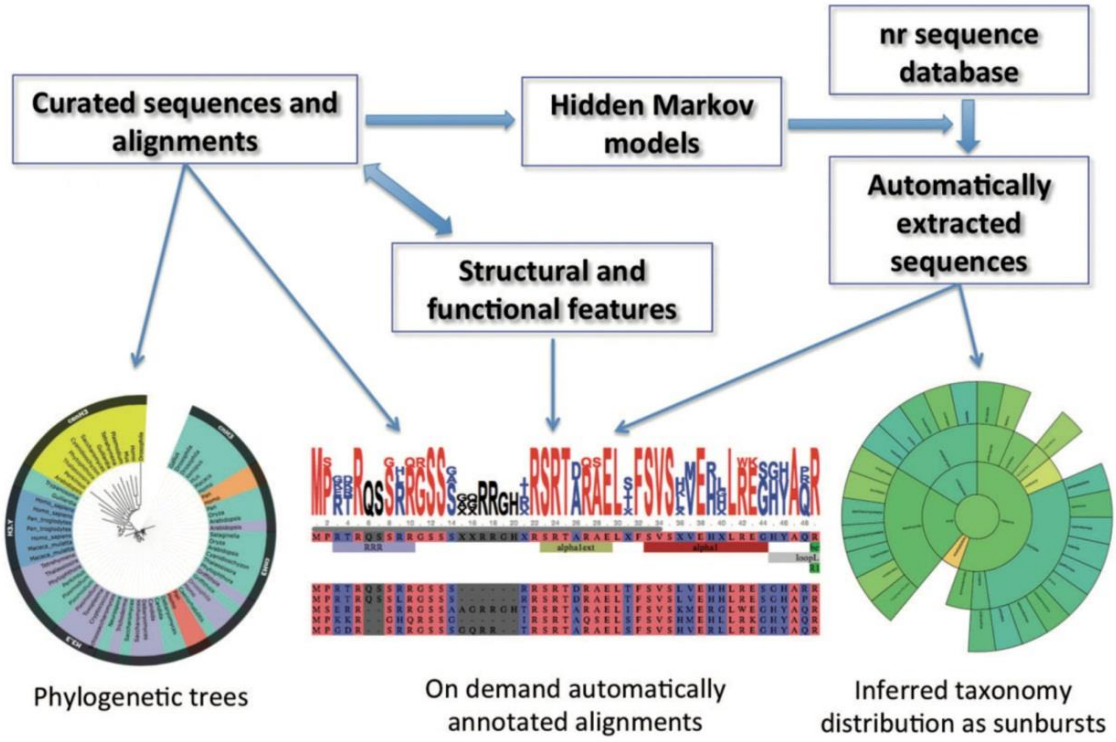

Рис. 6. Схематическое представление структуры HistoneDB 2.0 (Draizen et al., 2016).

Курируемый набор по (Draizen et al., 2016), создавался следующим образом. Последовательности гистоновых вариантов для каждого типа гистонов (H1, H2A, H2B, H3, H4) были собраны и проименованы согласно созданной в (Talbert and Henikoff, 2012) классификации. Эти последовательности были дополнены набором канонических гистонов, выровнены с помощью MUSCLE (Edgar, 2004) и дополнительно проверены вручную, чтобы убедиться в отсутствии вставок или делеций. В итоге полученный (Draizen et al., 2016) курируемый набор содержит гистоновые последовательности, относящиеся к 30 различным группам, которые

представляют собой наиболее известные гистоновые варианты и канонические гистоны.

Автоматически сгенерированный набор был создан авторами (Draizen et al., 2016) на основе курируемого с помощью скрытых марковских моделей (HMM).

4. Анализ последовательностей на основе скрытых марковских моделей.

Скрытые марковские модели (HMM) имеют ряд преимуществ перед стандартными методами работы с биологическими последовательностями. HMM применяют статистический подход для оценки истинной частоты встречаемости аминокислотного остатка в данной позиции по его наблюдаемой частоте. В то же время стандартные (не статистические) методы используют непосредственно наблюдаемую частоту для присвоения оценки этому остатку. Поэтому HMM, полученная из 10-20 выровненных последовательностей, может быть эквивалентна стандартному профилю, созданному из 40-50 выровненных последовательностей (Eddy, 1998). Эта особенность HMM позволила авторам (Draizen et al., 2016) использовать сравнительно небольшое количество последовательностей в курируемом наборе. По этой причине, в условиях обработки большого количества последовательностей с неопределенным статусом гомологии, как авторы HistoneDB 2.0, так и мы для текущей работы предпочли анализ на основе скрытых марковских моделей стандартным методам.

Скрытая марковская модель – модель, являющаяся направленным ациклическим графом. Она состоит из ряда узлов, каждый из которых соответствует позиции в выравнивании. В простой линейной модели каждый узел может существовать только в состоянии соответствия (т.е. в модели есть узел для каждой позиции в последовательности, которая должна быть выровнена с моделью). Для каждого узла есть две вероятности. Первая – вероятность перехода от одного узла к другому (в простой модели путь линеен, вероятность перехода равна 1). Вторая вероятность – вероятность соответствия этому узлу

конкретного аминокислотного остатка. Схема простой линейной модели представлена на рисунке 7.

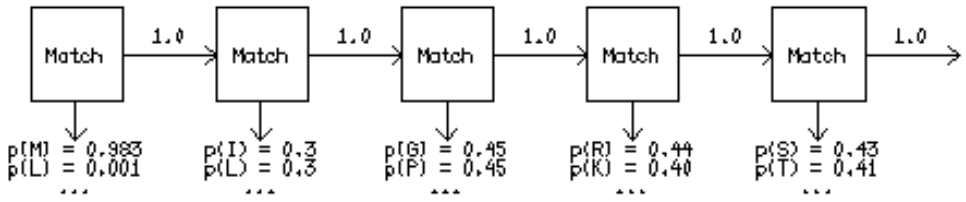

Рисунок 7. Схема простой линейной скрытой марковской модели (Eddy, 1998).

Для моделирования реальных последовательностей к модели добавляются еще два типа вероятностей: вероятности возникновения вставок и делеций. Первый тип возникает, когда последовательность содержит область, отсутствующую в модели (вставка в последовательность). Второй тип возникает, когда в модели есть область, которой нет в последовательности (делеция в последовательности). Для описания этих случаев каждый узел должен иметь уже три состояния: состояние соответствия, состояние вставки, состояние делеции. Схематическое изображение такой модели представлено на рисунке 8.

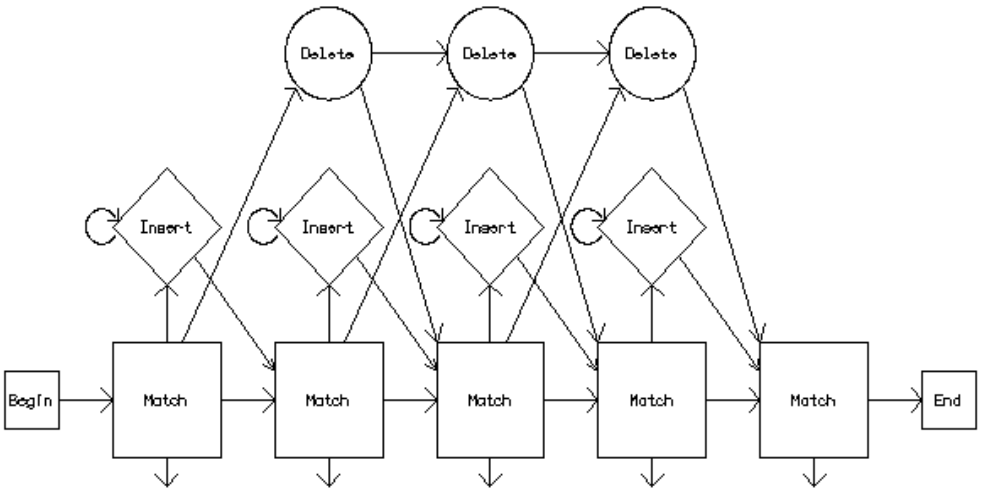

Рисунок 8. Схематическое представление нелинейной скрытой марковской модели (Eddy, 1998).

Если последовательность эквивалентна консенсусу исходного выравнивания, то путь через модель будет проходить от состояния соответствия

к состоянию соответствия линейным образом. Если последовательность содержит делецию относительно консенсуса, то путь проходит через одно или несколько состояний удаления перед переходом в следующее состояние соответствия; если последовательность содержит вставку относительно консенсуса, то путь проходит через состояние вставки между двумя состояниями соответствия.

Для анализа последовательностей на основе скрытых марковских моделей на практике чаще всего используется пакет HMMER (Eddy, Pearson, 2011). Это набор программ, который позволяет создавать профили на основе скрытых марковских моделей, производить чувствительный поиск по базам данных последовательностей и готовых профилей HMM. А также эффективно производить множественное выравнивание последовательностей.

Авторами базы данных (Draizen et al., 2016) были созданы модели HMM для каждой из 30 освещенных разновидностей гистонов на основе курируемого набора. В соответствии с этими моделями проверялись приходящие из не избыточных баз данных белковые последовательности. Все они классифицировались по той или иной модели варианта. Для каждой последовательности подсчитывались баллы HMMER, которые для регистрации того, что последовательность относится к ожидаемому варианту с 90% специфичностью, должны были превышать определенный порог. Специфичность выборки оценивалась по количеству истинных негативов (последовательностей, правильно классифицированных как не принадлежащих к варианту) и ложных позитивов (неверно предсказанные последовательности), найденных над каждым узлом HMMER. Специфичность была рассчитана как TN/(FP + TN). Желаемое значение порога оценки было получено из интерполированной обратной кривой графиков пороговых значений в зависимости от специфичности. В ряде случаев (например, гистон H2A.X) задавались также дополнительные условия наличия тех или иных конкретных мотивов.

5. Используемое программное обеспечение

Все статистические расчеты и анализ базы данных проводились с использованием высокоуровневого языка программирования Python 3.7 с подключением библиотек Pandas 1.0.3 (анализ данных), Scipy 1.2.3, Numpy 1.18.4 (работа с числами и векторами), Matplotlib 3.2.1 с приложением Seaborn 0.1.10 (визуализация результатов). Множественные выравнивания последовательностей строились в программе MUSCLE 3.8.31 (см. выше) с числом итераций, равным 10. Построение деревьев осуществлялось в программе MEGA 10.1.8 на основе алгоритма Neighbor joining (bootstrap = 500).

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

1. Распределение гистоновых вариантов по таксонам.

1.1 Распределение гистоновых вариантов по таксонам условного старшего уровня

На первом этапе работы мы получили распределение гистоновых вариантов по таксонам условного старшего уровня. Представленность в базе данных отображена на рисунке 9. Преобладают последовательности, полученные от представителей таксона Metazoa (61,27%). Относительно большие доли принадлежат Fungi (14,27%) и Plantae (19,61%). Слабее представлены организмы из группы SAR (2,91%). Количество последовательностей Discoba и Amoebozoa крайне мало в сравнении с остальными группами, поэтому в дальнейшем они не будут рассматриваться в настоящей работе.

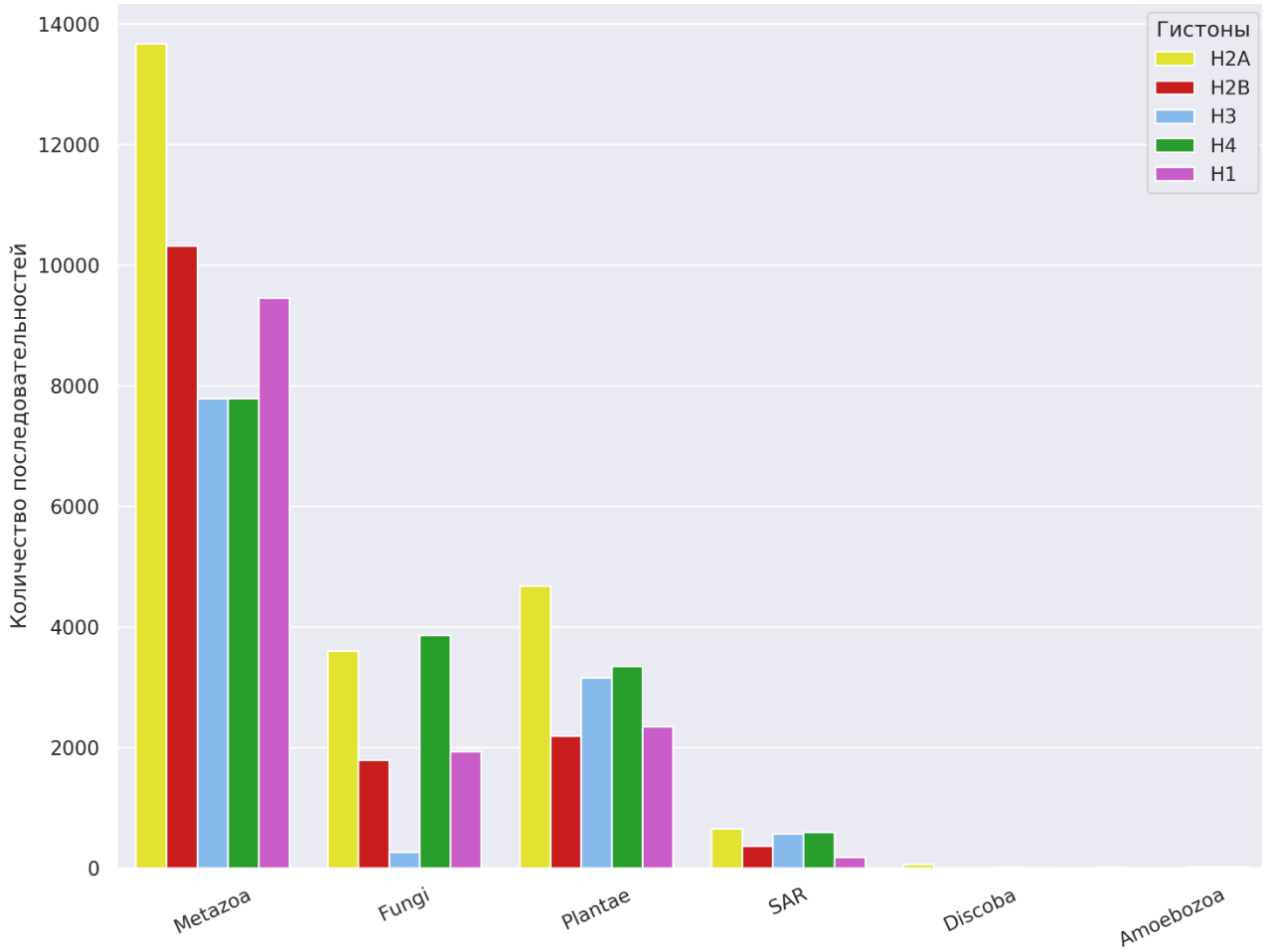

Рисунок 9. Распределение последовательностей гистоновых вариантов по таксонам условного старшего уровня.

Точное количество последовательностей каждого варианта для четырех групп старшего уровня представлено в таблице 4. Отдельно указано количество последовательностей из курируемого набора, использовавшихся при создании HMM-модели для каждого варианта. В силу специфики работы модели (таксономия изначально присвоена каждой последовательности, а модель относит её к тому или иному варианту) ошибки возможны только по вертикали. Иначе говоря, модель не может ошибочно отнести последовательность одного таксона к другому, но может отнести её к другому варианту вместо истинного.

Вариант засчитывался нами как представленный у таксона при соответствии хотя бы одному из следующих критериев:

1) количество последовательностей данного варианта у таксона превышает 5% от всех последовательностей данного варианта;
2) количество последовательностей данного варианта у таксона превышает 5% от общего количества последовательностей данного типа у таксона.

После результаты проверялись вручную с помощью поиска информации в литературных источниках и сравнения с таксономией последовательностей из курируемого набора. Наблюдаемая картина в массе своей совпадает с ожиданиями. Для каждой группы выявлены уникальные варианты, список которых будет дан в следующих подразделах. Как и ожидалось, большая их часть была зарегистрирована у представителей таксона Metazoa, у организмов из групп Fungi и SAR - по одному уникальному варианту, что также совпало с ожиданиями.

Однако некоторые результаты оказались неожиданными. Так, модель отнесла ряд последовательностей гистоновых белков SAR к варианту H2A.W, который согласно имеющимся данным, должен быть специфичен для растений. Схожая картина наблюдается у варианта H1.0: помимо животных модель также зарегистрировала его у грибов, у которых ранее он не наблюдался. В распределении последовательностей вариантов гистона H3 также много неожиданных результатов. Так, модель почти не зарегистрировала у растений вариант cenH3 (также известный как центромерный белок CENP-A), у грибов -

канонический H3 и H3.3, у SAR – канонический H2B, хотя данные варианты точно имеются у перечисленных таксонов.

Таблица 4. Распределение последовательностей гистоновых вариантов по таксонам условного старшего уровня. Красным шрифтом выделены варианты или канонические гистоны, распределение для которых расходится с ожидаемым.

Тип	variant_id	Metazoa	Fungi	Plantae	SAR	Curated set
H2A	canonical_H2A	7029	287	915	192	37
	macroH2A	2397	7	3	17	10
	H2A.1	1281	0	0	0	2
	H2A.B	277	1	0	0	15
	H2A.L	254	1	1	0	17
	H2A.P	139	0	0	0	11
	H2A.W	1	4	1927	111	9
	H2A.X	555	1474	356	36	23
	H2A.Z	1742	1826	1478	302	26
H2B	canonical_H2B	9077	1790	2186	4	29
	H2B.1	644	0	0	0	3
	H2B.W	387	0	0	0	6
	H2B.Z	8	5	0	358	2
	sperm_H2B	80	0	0	0	5
	subH2B	122	0	0	0	11
H3	canonical_H3	1210	8	2038	431	28
	cenH3	54	248	17	107	14
	H3.3	44	5	1091	33	15
	H3.5	2391	0	4	0	2
	H3.Y	133	0	1	0	8
	TS_H3.4	3955	2	4	0	2
H4	canonical_H4	7789	3865	3350	590	14
H1	generic_H1	6621	864	2344	161	18
	TS_H1.6	744	0	0	0	8
	TS_H1.7	199	7	0	5	2
	TS_H1.9	148	0	0	0	4
	OO_H1.8	490	0	0	0	2
	scH1	2	767	1	1	2
	H1.0	978	296	2	13	15
	H1.10	276	0	0	0	6

Для выяснения причины наблюдаемых результатов были произведены множественные выравнивания последовательностей из курируемого набора, на

основе которых в программа MEGA с использованием алгоритма Neighbor joining мы построили деревья, отражающие сходство последовательностей разных вариантов для спорных таксонов. В случае H2A.W мы также добавили курируемые последовательности Plantae и автоматически собранные последовательности предполагаемого варианта данного таксона. Упомянутые деревья не отражают каких-либо эволюционных взаимоотношений между вариантами, они призваны лишь продемонстрировать сходство последовательностей разных вариантов. Деревья будут представлены в соответствующих подразделах.

При рассмотрении таксонов уровня ниже старшего была выявлена неудовлетворительная точность работы марковской модели: она отнесла много сторонних последовательностей к вариантам с малым количеством курируемых последовательностей. Например, к вариантам H2B.Z и subH2B были причислены не только последовательности млекопитающих, но и в большом количестве последовательности далеких от них групп Metazoa, что, учитывая наличие всего 2 последовательностей каждого варианта в курируемом наборе, не является достоверным. В силу этого факта и обнаружения схожих проблем на уровне старших таксонов нами было принято решение не рассматривать группы рангом ниже типа или отдела.

1.2 Metazoa

Распределение последовательностей гистоновых вариантов по избранным группам условного младшего уровня внутри таксона Metazoa представлено на рисунке 10. Заметно преобладание типа Chordata (Хордовые), что, вероятнее всего, объясняется наибольшей популярностью этой группы организмов для секвенирования биологических последовательностей ввиду принадлежности к Хордовым человека и животных, имеющих большое научное, хозяйственное и рекреационное значение. На втором месте находится тип Arthropoda (Членистоногие), что можно объяснить чрезвычайно широкой распространенностью представителей этого таксона. Прочие группы

представлены слабо. Распределение последовательностей Metazoa по вариантам отображено на рисунке 11.

В ходе работы наибольшее количество уникальных вариантов было выявлено именно для Metazoa. Таковыми оказались macroH2A, H2A.1, H2A.B, H2A.L, H2A.P, H2B.1, H2B.W, sperm H2B, subH2B, H3.5, H3.Y, TS H3.4, TS H1.6, TS H1.7, TS H1.9, OO H1.8, H1.10. Вариант H1.0, вопреки ожиданиям, был обнаружен также у грибов, что будет подробнее рассмотрено в соответствующем подразделе. Наиболее спорные результаты касаются гистона H3. HMM отнесла к H3.3 слишком малое количество гистоновых последовательностей, выделенных из Metazoa, несмотря на то, что в курируемом наборе для указанного варианта имелись последовательности Metazoa. Возможно, это объясняется ошибкой скрытой марковской модели: она могла отнести к H3.3 последовательности, в действительности принадлежащие к другому варианту. Для проверки данного предположения мы построили множественные выравнивания курируемых последовательностей гистона H3 животных, на основе которых создали дерево,

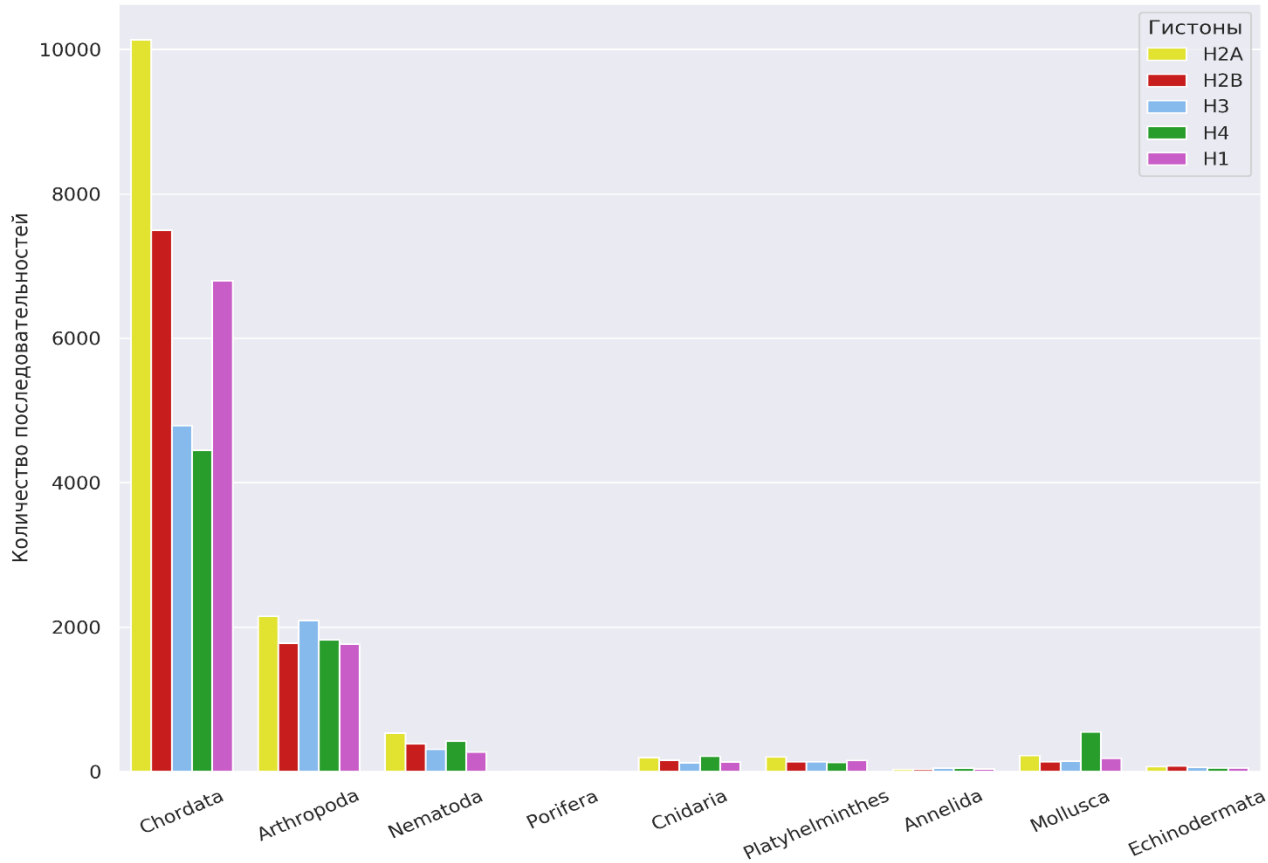

Рисунок 10. Распределение последовательностей гистоновых вариантов по избранным группам условного младшего уровня внутри таксона Metazoa.

отображающее сходства разных вариантов. Оно представлено на рисунке 12. По нему видна взаимная близость вариантов H3.3, H3.Y и H3.5. Можно предположить, что модель отнесла именно к двум последним вариантам последовательности H3.3. Чтобы проверить данное предположение, мы отобрали некоторые курируемые последовательности H3.3, H3.Y и автоматически собранные H3.Y, которые предположительно являются H3.3. Множественное выравнивание данных последовательностей можно видеть на рисунке 13. Заметно, что курируемые последовательности H3.3 и автоматически собранные H3.Y (из конкретного примера) проявляют между собой гораздо большую идентичность, чем с курируемыми H3.Y. Следует особенно отметить отсутствие у автоматически собранных H3.Y характерного Y-мотива, который и отличает указанный вариант от H3.3. Аналогично было проведено выравнивание с последовательностями гистона H3.5, в результате чего также были обнаружены автоматически извлеченные последовательности, отнесенные моделью к данному варианту, но на самом деле гораздо более близкие к H3.3. Все это говорит о значительной неточности работы модели при различении вариантов H3 у Metazoa.

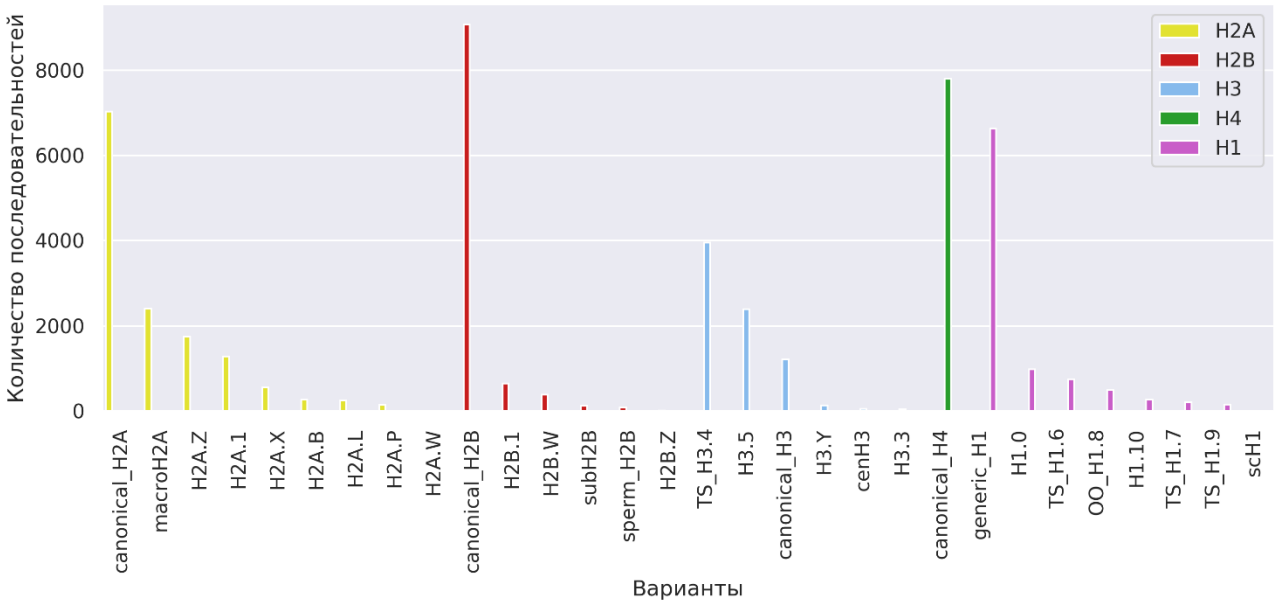

Рисунок 11. Распределение последовательностей по вариантам для таксона Metazoa.

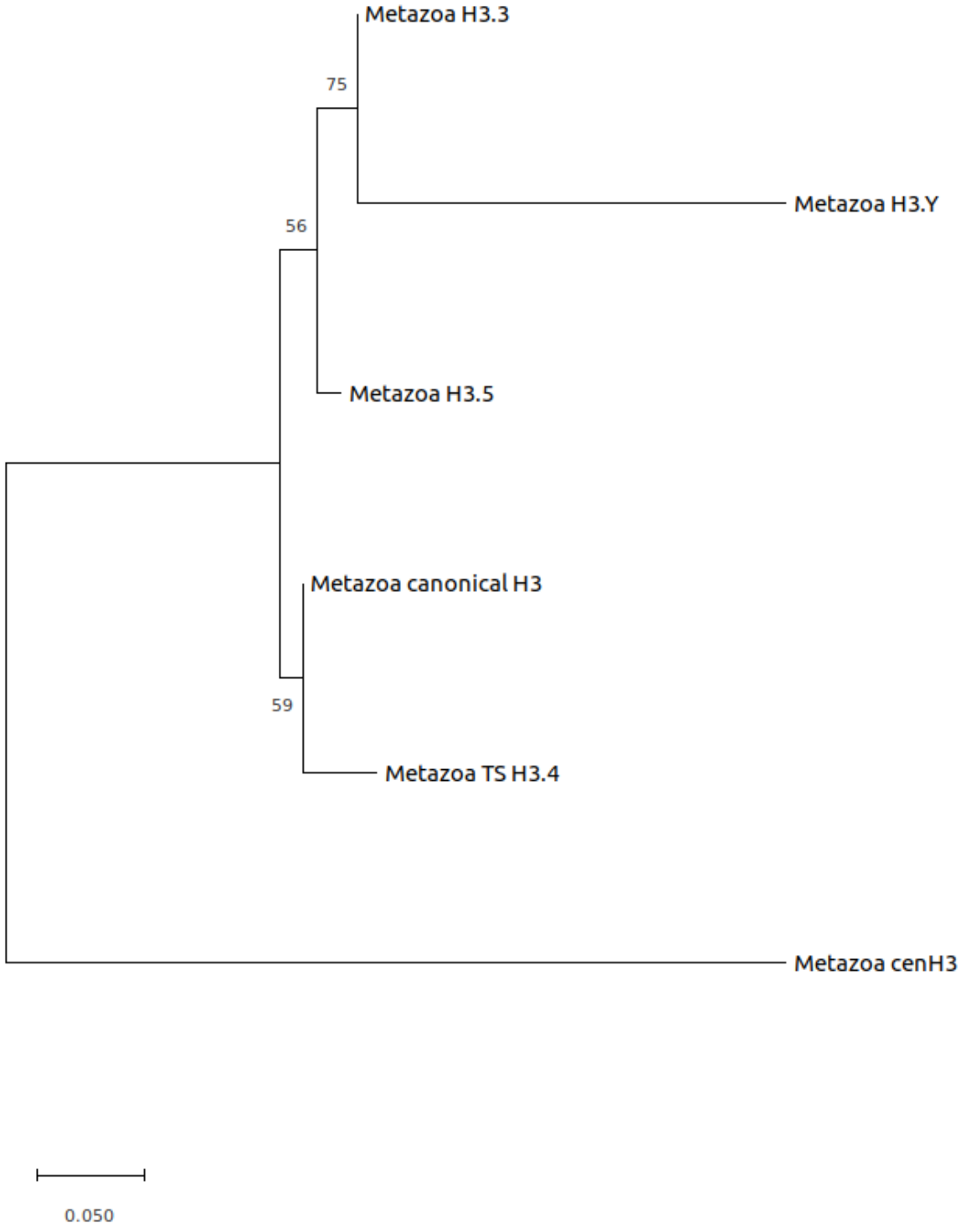

Рисунок 12. Филогенетическое дерево, построенное на основе курируемых последовательностей и отражающее сходство последовательностей разных вариантов гистона H3, обнаруженных у представителей таксона Metazoa.

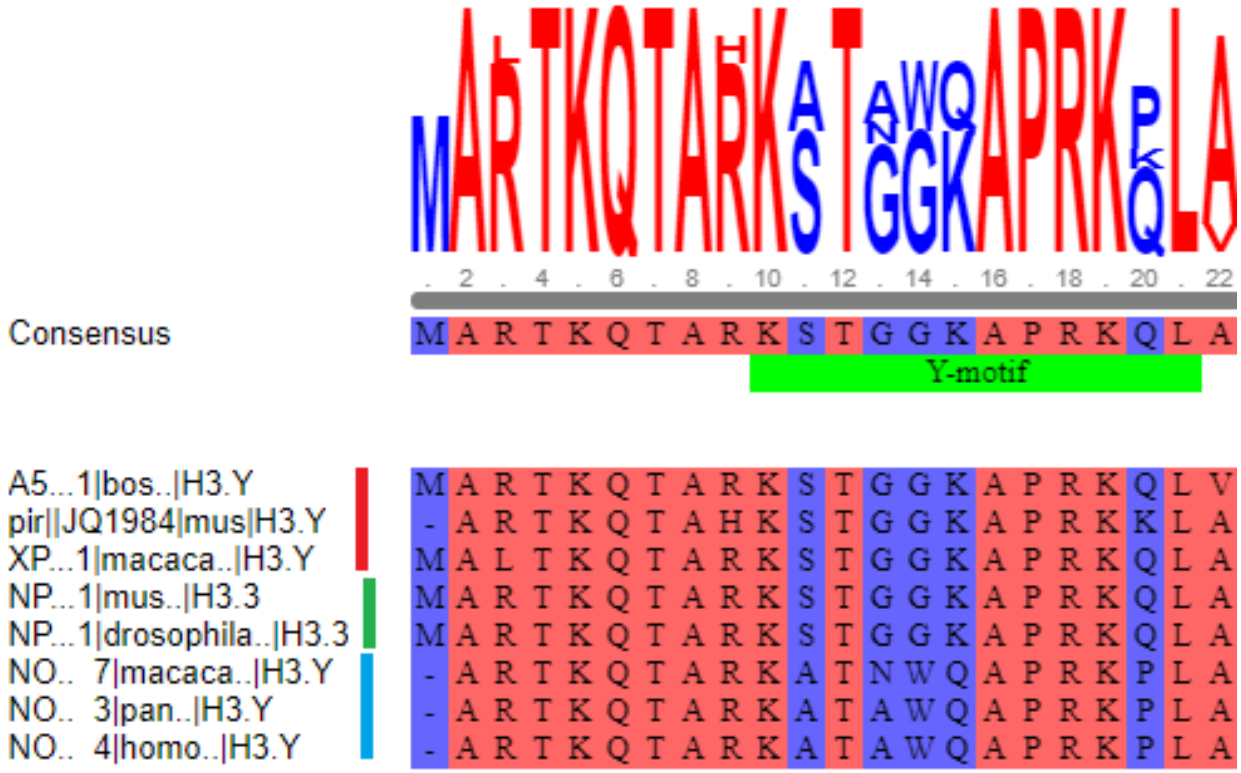

Рисунок 13. Множественное выравнивание избранных последовательностей H3.3 и H3.Y. Зеленой линией выделены курируемые последовательности H3.3, голубой - курируемые последовательности H3.Y, красной – примеры автоматически собранных последовательностей H3.Y. Синим маркером выделены сайты с 50% идентичностью, красным – с 80% идентичностью. Y-motif – особый мотив, отличающий H3.Y от H3.3.

1.3 Fungi

У таксона Fungi преобладают последовательности представителей высших грибов: аскомицетов (Ascomycota), базидиомицетов (Basidiomycota) (рисунок 14). Низшие грибы представлены слабо. Распределение последовательностей по вариантам для Fungi представлено на рисунке 15. Единственным вариантом для этого таксона условного старшего уровня является вариант гистона H1 - scH1, изначально обнаруженной у грибов рода Saccharomyces.

Однако результаты распределения для гистона H3 резко разошлись с ожидаемыми. Так, модель классифицировала крайне малое количество последовательностей Fungi, как принадлежащих к каноническому H3 и H3.3. При этом единственным вариантом H3 у грибов, который преодолел 5% порог

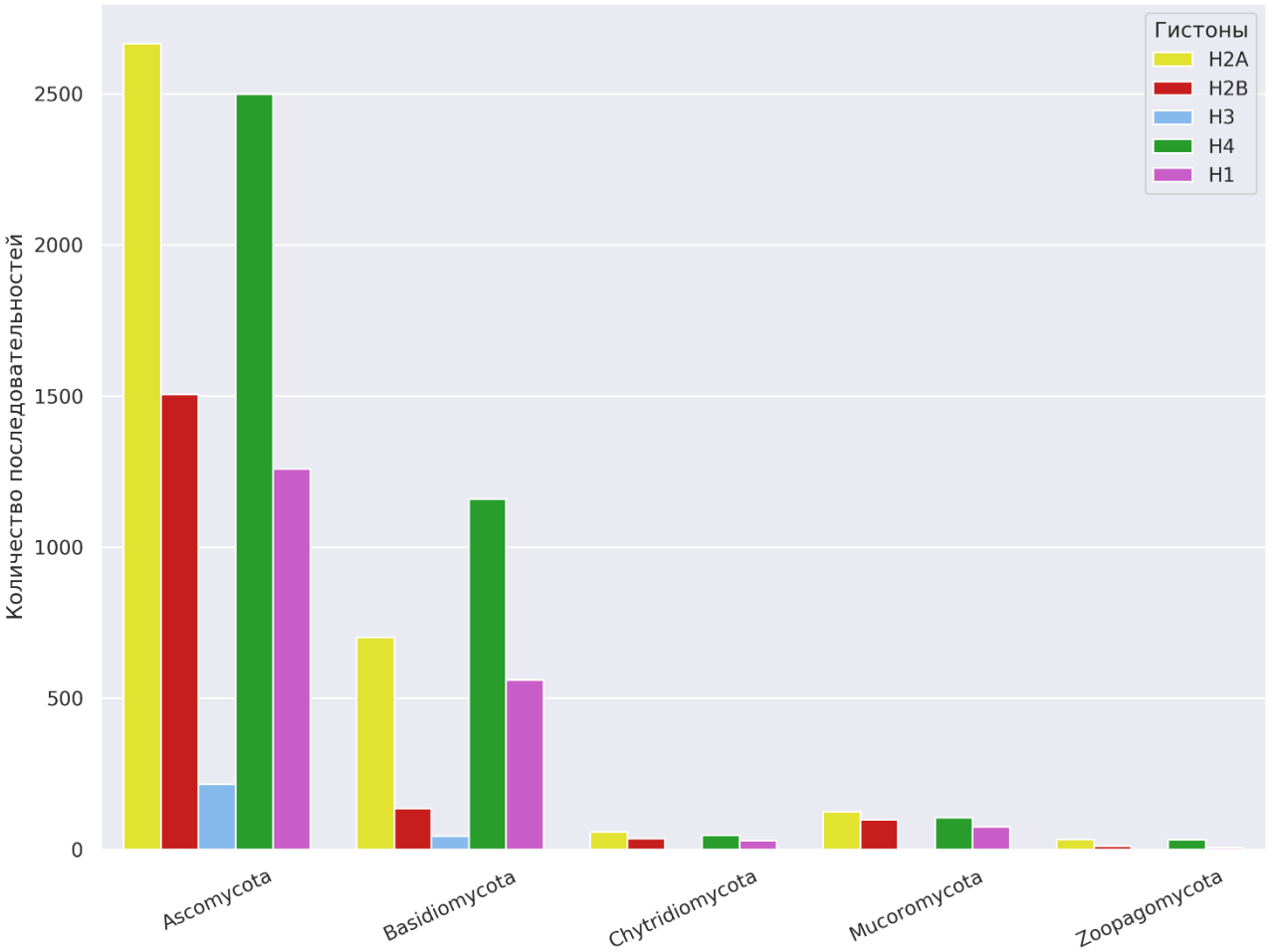

Рисунок 14. Распределение последовательностей гистоновых вариантов по избранным группам условного младшего уровня внутри таксона Fungi.

(см. подраздел 1), является центромерный гистон cenH3. Модель, вероятнее всего, не могла ошибочно отнести к этому варианту последовательности канонического H3 и H3.3, так как идентичность между cenH3 и остальными вариантами H3 достаточно низка. Это значит, что последовательности, принадлежащие Fungi, не прошли проверку модели и не были добавлены в базу данных. Вероятно, это связано с относительно высокой уникальностью грибных H3.3 и канонического H3 по сравнению с указанными вариантами других таксонов, а также с их высокой идентичностью относительно друг друга (различаются всего на несколько аминокислот).

Расходится с ожидаемой также картина распределения по вариантам H1. А именно: гистон H1.0 помимо животных также был обнаружен у грибов. Вероятнее всего модель причислила к данному варианту последовательности

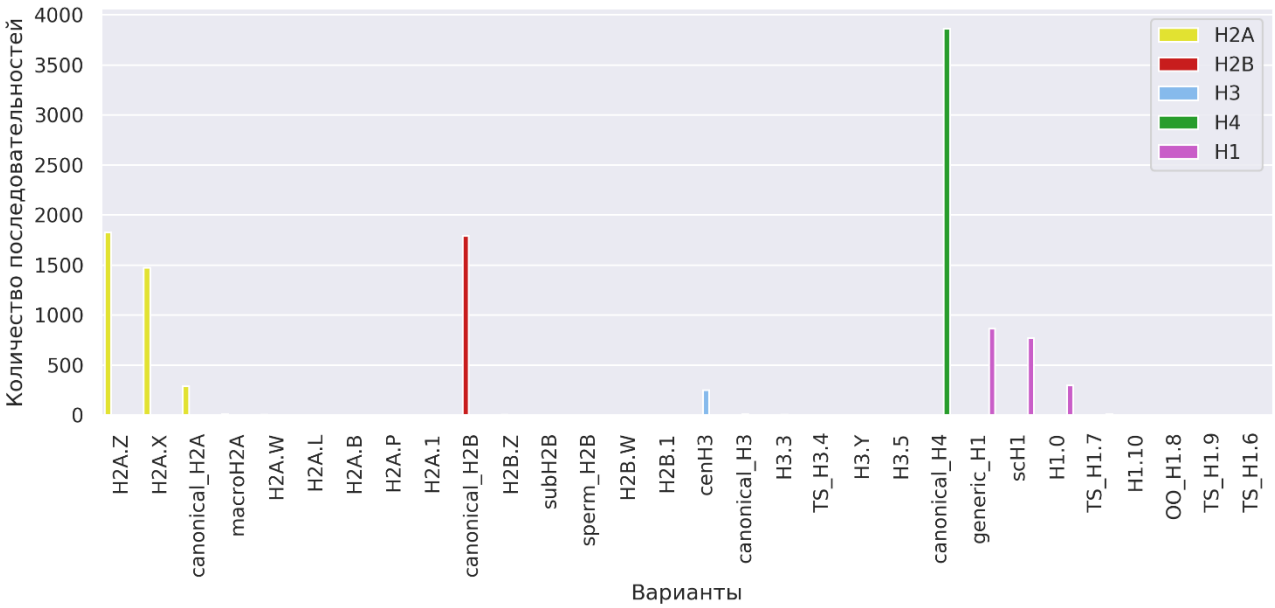

Рисунок 15. Распределение последовательностей по вариантам для таксона Fungi.

грибов, в действительности, принадлежащие к generic H1 и scH1. Так как в курируемом наборе для грибов был всего один вариант H1 (scH1), построение полноценного дерева не представлялось возможным. Поэтому для проверки предположения мы выясняли степень идентичности курируемых последовательностей scH1, generic H1 и H1.0 и собранных автоматически последовательностей H1.0 грибов. Последние действительно показали высокую идентичность с животным H1.0 и крайне низкую с scH1. Это может говорить о действительном наличии H1.0 у таксона Fungi. Более подробное изучение данного вопроса является перспективным для будущих исследований.

1.4 Plantae

Среди Plantae в базе данных подавляющее превосходство у сосудистых растений (Tracheophyta). Слабее представлены зеленые водоросли (Chlorophyta). Прочие группы представлены крайне слабо (рисунок 16). Распределение последовательностей по вариантам для Plantae демонстрируется на рисунке 17.

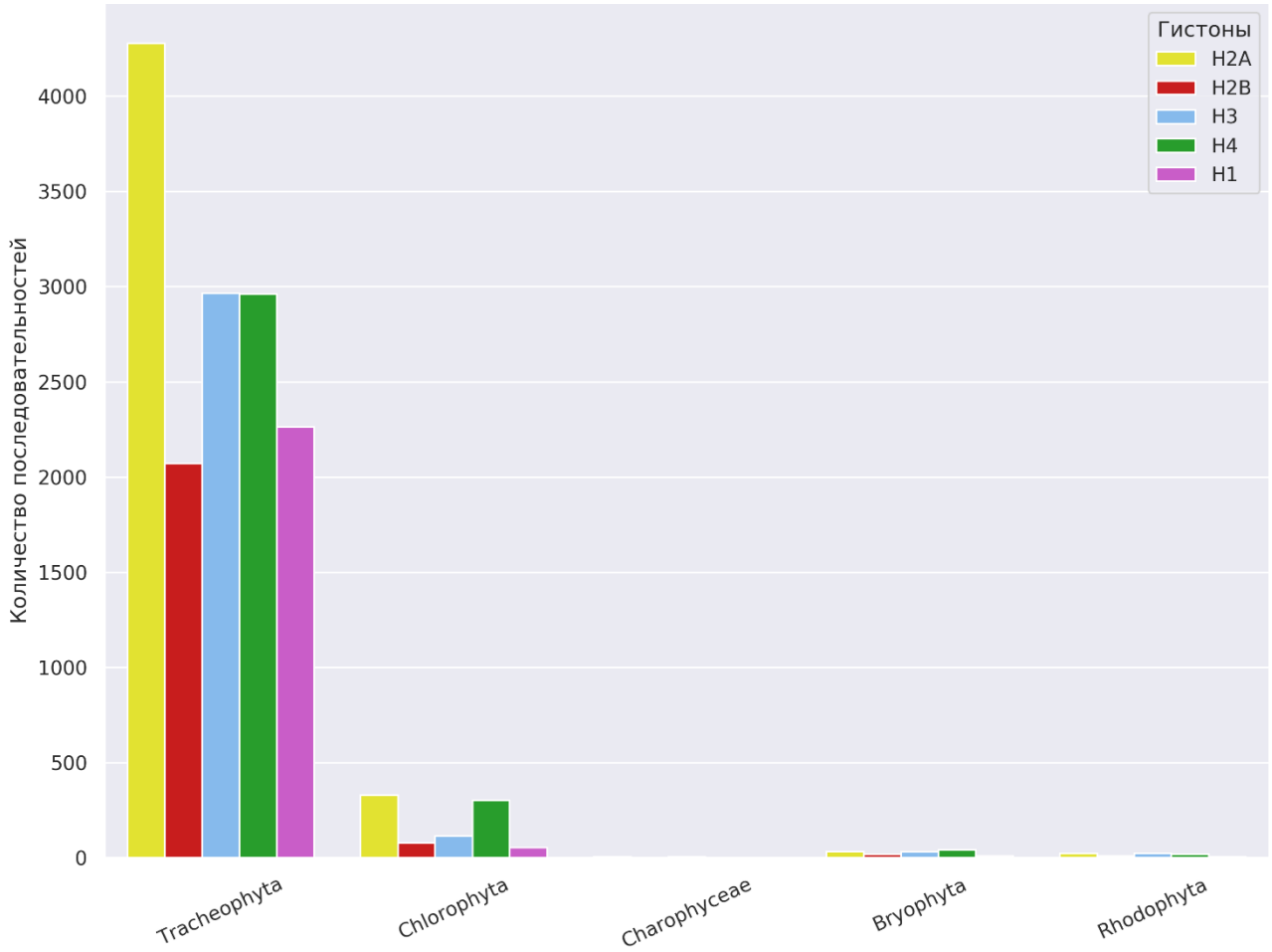

Рисунок 16. Распределение последовательностей гистоновых вариантов по избранным группам условного младшего уровня внутри таксона Plantae.

Вопреки ожиданиям уникальных для растений вариантов обнаружено не было. H2A.W, ожидавшийся как специфичный для Plantae был обнаружен моделью также у SAR. Подробнее этот вопрос будет рассмотрен в следующем подразделе. Также модель классифицировала как cenH3 крайне малое количество последовательностей, принадлежащих Plantae. Это нельзя объяснить тем, что модель отнесла последовательности растительного cenH3 к другим вариантам данного типа ввиду значительной уникальности cenH3. Вероятнее всего, это объясняется малой идентичностью последовательностей cenH3 между собой, в результате чего многие последовательности просто не проходят проверку на соответствие.

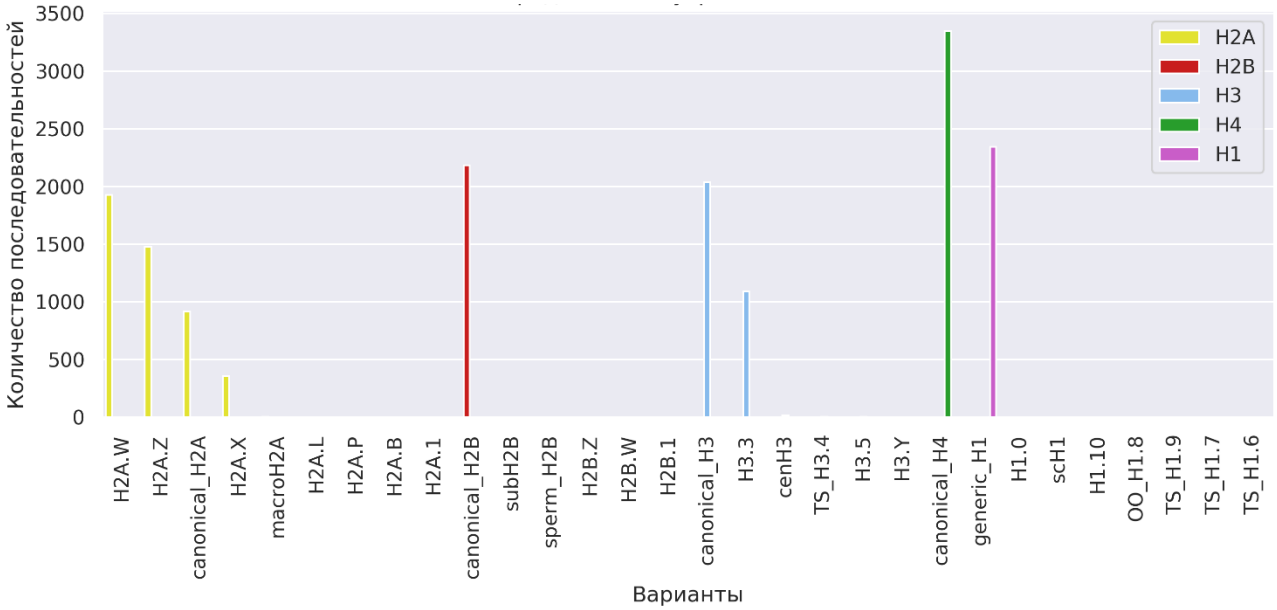

Рисунок 17. Распределение последовательностей по вариантам для таксона Plantae.

1.5 SAR

Внутри таксона SAR в базе данных преобладают группы Oomycota и Apicomplexa. Менее представлены Ochrophyta и Ciliophora. Крайне малое количество последовательностей относится к Dinophyta (в NCBI Tacsonomy синоним – Dinophyceae), Cercozoa и Retaria (рисунок 18).

Распределение по вариантам для данного таксона условного старшего уровня представлено на рисунке 19. Уникальным для SAR вариантом является H2B.Z, что совпало с ожиданиями. Но при этом модель почти не обнаружила у SAR канонического H2B. Так как в курируемом наборе всего одна последовательность канонического H2B SAR, можно предположить, что модель отнесла автоматически собранные последовательности типа H2B SAR к H2B.Z. Для проверки этого утверждения мы произвели множественное выравнивание курируемых последовательностей H2B.Z и канонического H2B SAR, автоматически собранных последовательностей H2B.Z, а также курируемых последовательностей канонического H2B других таксонов. В результате все последовательности SAR (курируемые и собранные автоматически, канонического H2B и H2B.Z) показали между собой более чем 90%

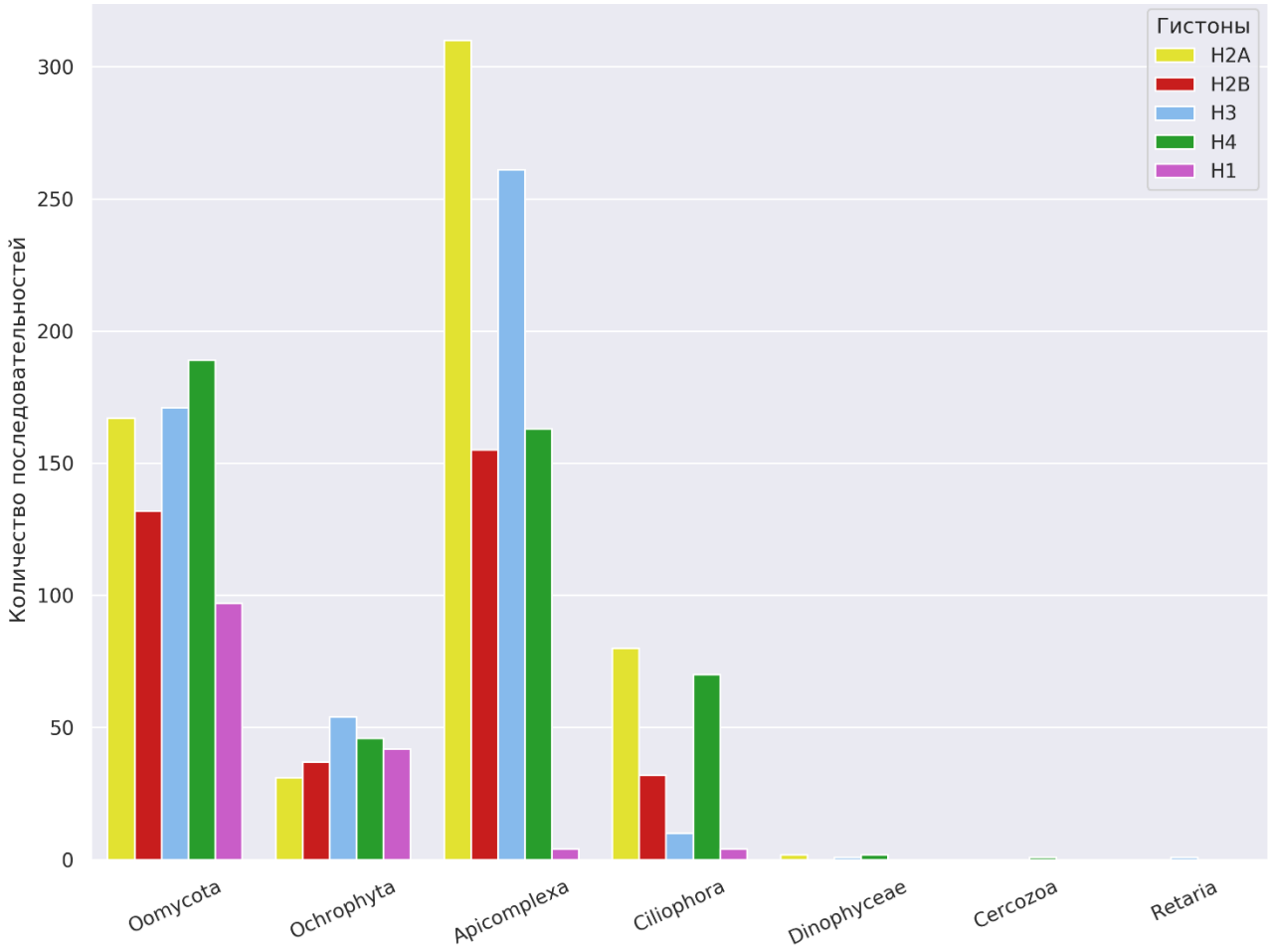

Рисунок 18. Распределение последовательностей гистоновых вариантов по избранным группам условного младшего уровня внутри таксона SAR.

идентичность, чего нельзя сказать об идентичности с остальными таксонами условного старшего уровня. Таким образом, наблюдаемое распределение действительно может объясняться тем, что модель классифицировала ряд последовательностей канонического H2B SAR как H2B.Z.

Интересным результатом является также обнаружение моделью у SAR варианта H2A.W, который должен быть специфичным для растений. Для проверки того, действительно ли указанный вариант имеется у SAR или имела место быть некорректная работа модели, мы отобрали курируемые последовательности всех вариантов H2A SAR, добавили к ним курируемые последовательности растительного H2B.W и некоторые автоматически собранные последовательности предполагаемого H2A.W SAR. Внутри каждой группы были произведены множественные выравнивания и получены консенсусы, на основе

которых было построено дерево, отражающее взаимную близость указанных вариантов (рисунок 20). Из него можно заметить, что последовательности SAR, предсказанные моделью как относящиеся к H2A.W, на самом деле значительно

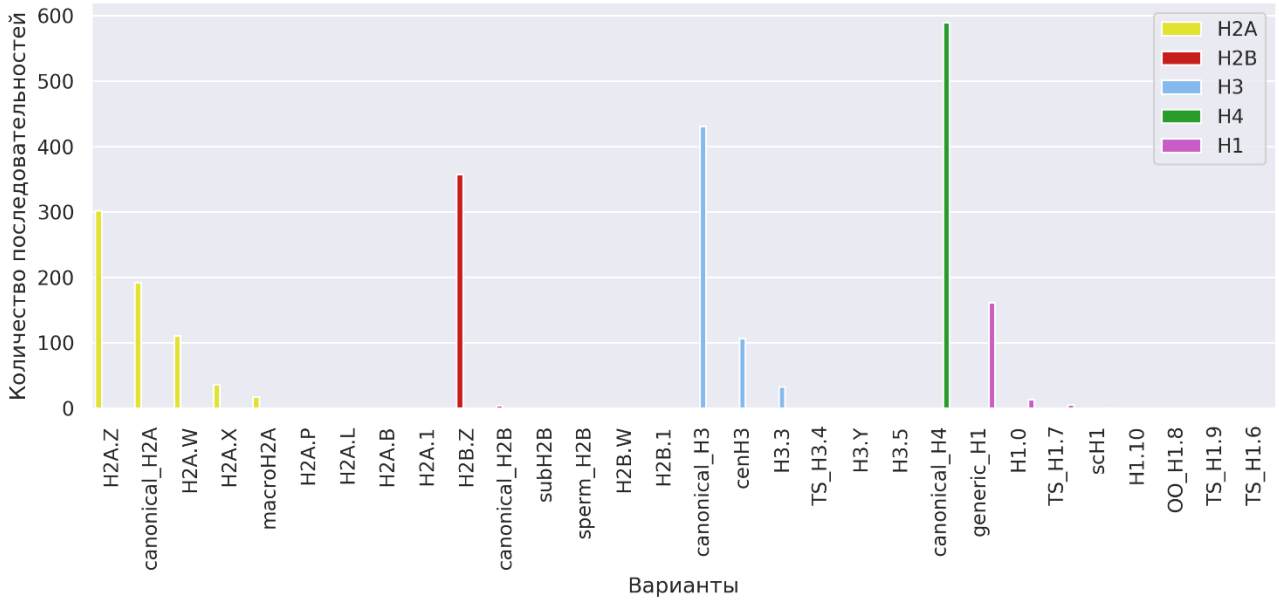

Рисунок 19. Распределение последовательностей по вариантам для таксона SAR.

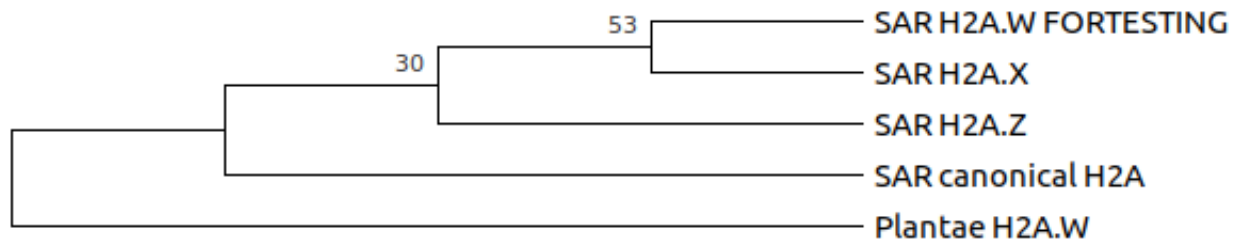

Рисунок 20. Филогенетическое дерево, отражающее сходство последовательностей разных вариантов гистона H2A, обнаруженных у представителей таксона SAR. Для уточнения корректности работы HMM в отношении варианта H2A.W рассмотрен также растительный H2A.W. H2A.W FORTESTING – предполагаемый H2A.W SAR, выявленный HMM.

ближе к другим вариантам данного типа у SAR, чем к последовательностям растительного H2A.W. Данный результат также подтверждается отсутствием в последовательностях предполагаемого H2A.W SAR SPKK-мотива – характерного консервативного С-концевого мотива, специфичного для H2A.W (рисунок 21). Все это говорит о том, что при отборе последовательностей моделью произошла ошибка: к варианту H2A.W были отнесены

последовательности других вариантов SAR. Иначе говоря, в действительности, у SAR нет данного варианта, H2A.W специфичен для растений.

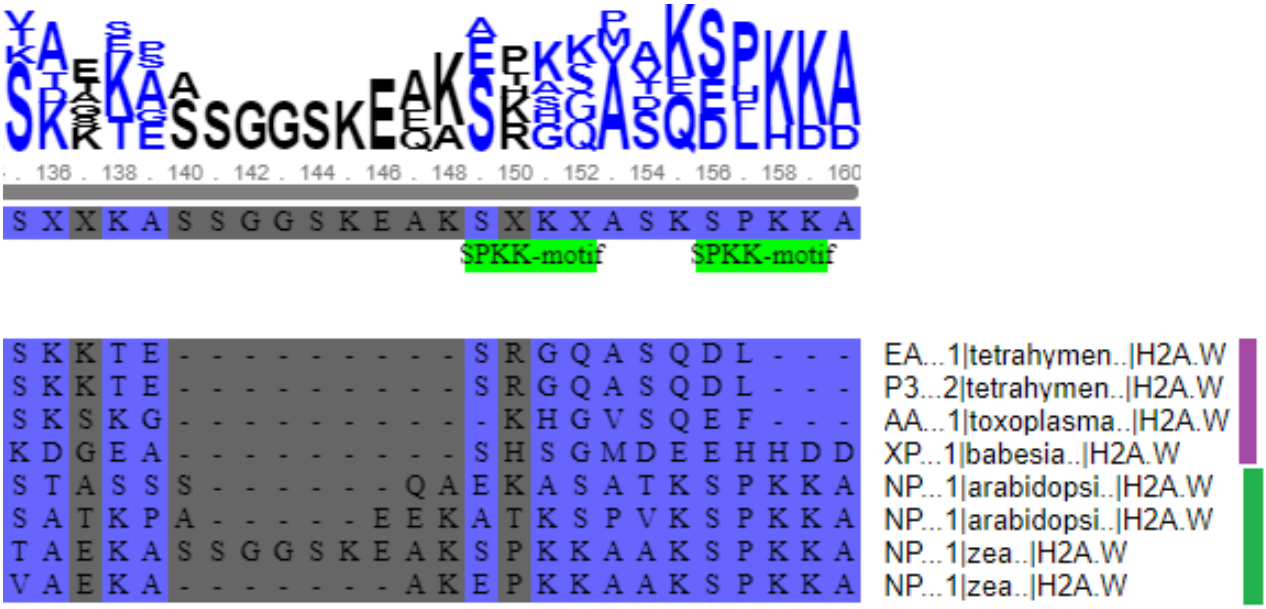

Рисунок 21. Изображение в области SPKK-мотива множественного выравнивания избранных курируемых последовательностей H2.A.W Plantae (зеленая линия) и автоматически отобранных последовательностей SAR, предположительно также относящихся к варианту H2A.W (фиолетовая линия).

2. Распределение гистоновых вариантов по длинам

Вторая часть работы заключалась в выявлении распределения разных групп гистоновых последовательностей по длинам. Соответствующие статистические данные для суммарных типов гистонов, всех гистоновых вариантов и для типов четырех рассматриваемых таксонов условного старшего уровня содержатся в таблицах 5, 6 и 7. Диаграммы типа box plot, отражающие вышеуказанные данные для рассматриваемых вариантов, можно увидеть на рисунке 22.

Из таблицы 5 видно, что наиболее короткие последовательности характерны для гистона H4 - 103 аминокислотных остатка. Последовательности гистонов H3 и H4 наиболее консервативны по длине: интерквартиль (разность третьего и первого квартилей) равен 0. Гистоны H2A и H2B менее консервативны по длине: интерквартиль равен 15 для H2A и 13 для H2B. Это интересное наблюдение, ведь

в структуре нуклеосомы гистоны объединены в димеры именно так: H2A объединен с H2B, а H3 с H4. Наиболее вариабельным по длине является линкерный гистон H1. Этот результат совпадает с ожидаемым: согласно имеющимся данным H1 не гомологичен к коровым гистонам и эволюционировал независимо.

В таблице 6 можно видеть более подробные данные о распределении последовательностей каждого варианта. В целом можно отметить значительную стабильность в длине для гистоновых последовательностей: интерквартиль у большинства вариантов не превышает 10 аминокислотных остатков. В таблице есть примеры более (H2A.W, H2B.W) и менее крупных (H2A.B, канонические H4). Центромерный H3 демонстрирует наибольшую изменчивость в длине последовательностей среди всех вариантов коровых гистонов (интерквартиль равен 90), что подтверждает данные о неоднородности этого варианта и противоречивых свойствах его последовательностей. Варианты H1 демонстрируют самую высокую вариабельность по длине. Самый маленький из них – TS H1.9, самый большой – OO H1.8.

Таблица 7 демонстрирует распределение по длинам последовательностей гистонов пяти типов для таксонов условного старшего уровня. Гистоны H4 и H3 показывают тенденцию к стабильности длины независимо от таксона. Расхождение для гистона H3 Fungi связано с крайне малым количеством автоматически собранных последовательностей этого типа и преобладанием последовательностей центромерного H3, который является наиболее изменчивым вариантом коровых гистонов; этот вопрос был подробно рассмотрен выше.

Гистоны H2A и H2B показывают более значительную вариабельность по длинам между рассматриваемыми таксонами. Наиболее короткие последовательности H2A обнаружены у Metazoa. Самые короткие H2B – у SAR. Для них же характерны более короткие, чем у других таксонов, последовательности линкерного H1.

Таблица 5. Распределение по длинам последовательностей для типов гистонов.

Тип	N пос-тей	Min	1Q	Med	3Q	Max
H2A	23000	72	128	133	143	2344
H2B	14895	66	124	126	137	3615
H3	11924	97	136	136	136	1449
H4	15978	79	103	103	103	2455
H1	14151	40	198	223	288	3229

Таблица 6. Распределение по длинам последовательностей для канонических гистонов и гистоновых вариантов.

Тип	Вариант	N пос-тей	Min	1Q	Med	3Q	Max
H2A	Canonical H2A	8482	107	125	128	130	2344
	macroH2A	2436	111	368	370	372	572
	H2A.1	1328	128	130	130	130	187
	H2A.B	285	72	113	115	119	833
	H2A.L	257	83	110	117	118	921
	H2A.P	141	72	112	117	117	234
	H2A.W	2092	111	143	149	153	886
	H2A.X	2487	122	132	134	139	463
	H2A.Z	5492	82	129	136	141	1081
H2B	Canonical H2B	13275	91	124	126	137	3615
	H2B.1	658	125	126	126	127	207
	H2B.W	389	89	142	163	175	307
	H2B.Z	371	95	119	123	123	673
	sperm_H2B	80	103	121	123	140	297
	subH2B	122	66	122	122	123	204
H3	Canonical H3	3746	133	136	136	136	1449
	cenH3	429	97	139	179	229	448
	H3.3	1180	136	136	136	136	152
	H3.5	2419	133	136	136	136	372
	H3.Y	136	131	136	136	146	766
	TS H3.4	4014	134	136	136	136	509
H4	Canonical H4	15978	79	103	103	103	2455
H1	Generic H1	10123	40	202	229	323	3229
	TS H1.6	753	150	212	219	221	393
	TS H1.7	227	75	255	271	331	2792
	TS H1.9	148	94	167	173	180	1045
	OO H1.8	494	69	276	327	447	1444
	scH1	784	55	199	225	258	2596
	H1.0	1344	54	185	194	206	1983
	H1.10	278	45	194	213	218	1520

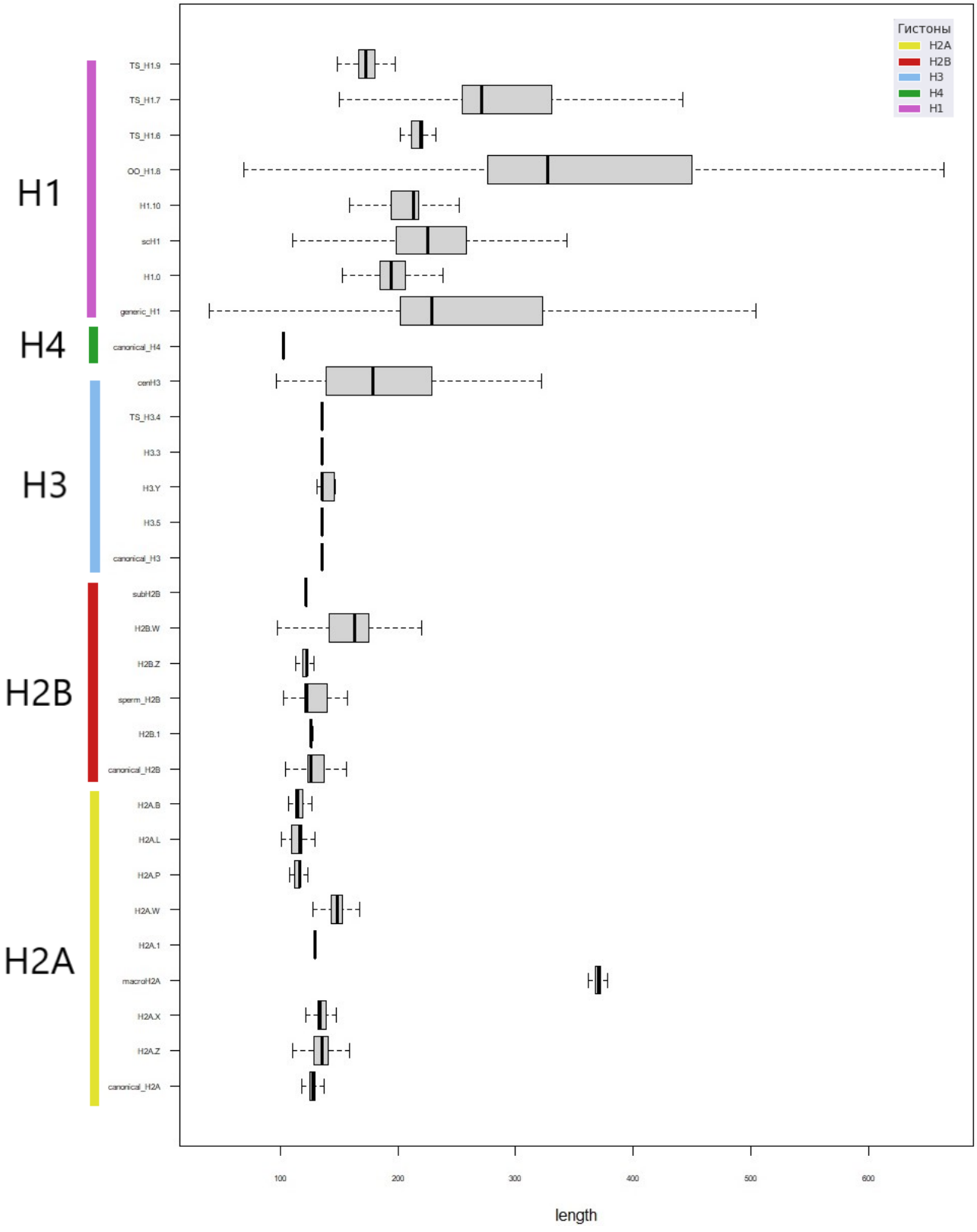

Рисунок 22. Диаграммы типа “Box plot”, отражающие распределение последовательностей рассматриваемых гистоновых вариантов по длинам. Выбросы из статистики исключены из рассмотрения (выбросами считались все значения, не включенные в интервал (Q1 – 1,5IQ; Q3 + 1,5IQ); Q1, Q3 – первый и третий квартили, IQ – интерквартиль (Q3 – Q1)

Таблица 7. Распределение по длинам последовательностей для таксонов условного старшего уровня.

Тип	Таксон	N пос-тей	Min	1Q	Med	3Q	Max
H2A	Metazoa	13675	72	127	129	142	2344
	Fungi	3600	106	133	135	139	1217
	Plantae	4680	82	134	138	149	1593
	SAR	658	99	133	139	155	890
H2B	Metazoa	10318	66	124	126	126	1713
	Fungi	1795	100	131	137	140	410
	Plantae	2186	100	138	143	148	695
	SAR	362	95	119	123	123	254
H3	Metazoa	7787	129	136	136	136	1449
	Fungi	263	120	172	229	229	424
	Plantae	3155	127	136	136	136	437
	SAR	571	97	136	136	136	448
H4	Metazoa	7789	79	103	103	103	1041
	Fungi	3865	84	103	103	103	2455
	Plantae	3350	82	103	103	103	1713
	SAR	590	82	103	103	103	464
H1	Metazoa	9458	40	199	221	332	2297
	Fungi	1934	55	196	221	253	2596
	Plantae	2347	64	196	253	295	1615
	SAR	180	47	165	189	243	3229

ВЫВОДЫ

1. Проведен количественный анализ встречаемости последовательностей белков гистонов в базе данных GenBank. Определено распределение последовательностей по таксонам условного старшего уровня: на перовом месте Metazoa (61,27%), на втором - Plantae (19,61%), на третьем – Fungi (14,27%), на четвертом SAR – (2,91%). Последовательностей прочих групп крайне мало.

2. Произведен анализ филетического распределения гистоновых вариантов по таксонам. Для каждого таксона условного старшего уровня из представленных в базе данных вариантов выделены уникальные: 17 для Metazoa, по одному для Fungi, Plantae и SAR.

3. У таксона Fungi обнаружены последовательности, схожие с H1.0 животных. Это может говорить о действительном наличии данного варианта у грибов. Проверка этого предположения является перспективной темой будущих исследований.

4. Система классификации последовательностей по вариантам на основе скрытых марковских моделей показала неточность работы при различении таксонов рангом ниже типа или отдела, а также при различении вариантов гистона H3 у Metazoa и Fungi. Наименьшая точность классификации наблюдалась у вариантов с малым количеством последовательностей в курируемом наборе. Это говорит о необходимости усовершенствования базы данных для более эффективного различения последовательностей близких таксонов и близких изоформ гистонов.

5. Выявлена значительная стабильность в длине гистоновых последовательностей. Гистоны H3 и H4 - наиболее консервативны по длине, которая вне зависимости от таксона составляет 136 аминокислотных остатков для H3 и 103 аминокислотных остатка для H4. H2A и H2B более вариабельны: характерная длина H2A 128 – 143 аминокислотных остатка, для H2B – 124 – 137 аминокислотных остатков. Наиболее крупным и вариабельным является линкерный гистон H1 (198 – 288 остатков).

6. Для группы SAR характерны наиболее короткие последовательности H2B (119 – 123 аминокислотных остатка) и H1 (165 – 243 аминокислотных остатка).

ЗАКЛЮЧЕНИЕ

В данной работе были проанализированы особенности филетического распределения гистоновых вариантов в эукариотических организмах четырех групп: Metazoa, Fungi, Plantae, SAR. В ряде случаев было произведено детальное сравнение последовательностей методом множественных выравниваний, на основе которых строились деревья, отражающие их сходства. Также были рассмотрены характерные свойства распределения последовательностей гистоновых вариантов по длинам.

СПИСОК ЛИТЕРАТУРЫ

1. Sina M Adl, Alastair G B Simpson, Mark A Farmer, Robert A Andersen, O Roger Anderson, John R Barta, Samuel S Bowser, Guy Brugerolle, Robert A Fensome, Suzanne Fredericq, Timothy Y James, Sergei Karpov, Paul Kugrens, John Krug, Christopher E Lane, Louise A Lewis, Jean Lodge, Denis H Lynn, David G Mann, Richard M McCourt, Leonel Mendoza, Ojvind Moestrup, Sharon E Mozley-Standridge, Thomas A Nerad, Carol A Shearer, Alexey V Smirnov, Frederick W Spiegel, Max F J R Taylor. The new higher level classification of eukaryotes with emphasis on the taxonomy of protists // The Journal of Eukaryotic Microbiology. 2005. 52 (5): 399–451.

2. Sina M Adl, Alastair G B Simpson, Christopher E Lane, Julius Lukeš, David Bass, Samuel S Bowser, Matthew W Brown, Fabien Burki, Micah Dunthorn, Vladimir Hampl, Aaron Heiss, Mona Hoppenrath, Enrique Lara, Line Le Gall, Denis H Lynn, Hilary McManus, Edward A D Mitchell, Sharon E Mozley-Stanridge, Laura W Parfrey, Jan Pawlowski, Sonja Rueckert, Laura Shadwick, Conrad L Schoch, Alexey Smirnov, Frederick W Spiegel. The revised classification of eukaryotes //The Journal of Eukaryotic Microbiology. 2012. 59 (5): 429–93.

3. Arents G., Burlingame R.W., Wang B.C., Love W.E., Moudrianakis E.N. The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix // Proc. Natl. Acad. Sci. 1991. 88(22): 10148–10152.

4. Arents G., Moudrianakis E.N. Topography of the histone octamer surface: repeating structural motifs utilized in the docking of nucleosomal DNA // Proc. Natl. Acad. Sci. 1993. 90(22): 10489–10493.

5. Barrero, M.J., Sese B., Martí M., Izpisua Belmonte J.C. Macro histone variants are critical for the differentiation of human pluripotent cells // J. Biol. Chem. 2013. 288(22): 16110–16116.

6. Barrero, M.J., Sese B., Kuebler B., Bilic J., Boue S., Martí M., Izpisua Belmonte J.C. Macrohistone Variants Preserve Cell Identity by Preventing the Gain

of H3K4me2 during Reprogramming to Pluripotency // Cell Rep. 2013a. 3(4): 1005–1011.

7. Baxevanis A.D., Landsman D. Histone Sequence Database: a compilation of highly-conserved nucleoprotein sequences. // Nucleic Acids Res. 1996. 24(1): 245–247.

8. Baxevanis A.D., Landsman D. Histone Sequence Database: new histone fold family members // Nucleic Acids Res. 1998. 26(1): 372–375.

9. Burki F. The eukaryotic tree of life from a global phylogenomic perspective // Cold Spring Harbor Perspectives in Biology. 2014. 6 (5)

10. Chadwick B.P., Willard H.F. Histone H2A variants and the inactive X chromosome: identification of a second macroH2A variant. // Hum. Mol. Genet. 2001. 10(10) 1101–1113.

11. Chadwick B.P., Willard H.F. Multiple spatially distinct types of facultative heterochromatin on the human inactive X chromosome // Proc. Natl. Acad. Sci. U. S. A. 2004. 101(50): 17450–17455.

12. Chakravarthy S., Gundimella S.K.Y., Caron C., Perche P.-Y., Pehrson J.R., Khochbin S., Luger K. Structural characterization of the histone variant macroH2A // Mol. Cell. Biol. 2005. 25(17): 7616–7624.

13. Crane-Robinson C., Dancy S.E., Bradbury E.M., Garel A., Kovacs A.M., Champagne M., Daune M. Structural studies of chicken erythrocyte histone H5. // Eur. J. Biochem. 1976. 67(2): 379–388.

14. DeLange R.J., Smith E.L. Histones: structure and function. // Annu. Rev. Biochem. 1971. 40, 279–314.

15. Draizen E.J., Shaytan A.K., Marino-Ramirez L., Talbert P.B., Landsman D., Panchenko A.R. HistoneDB 2.0: a histone database with variants--an integrated resource to explore histones and their variants. // Database (Oxford). 2016. 2016.

16. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput // Nucleic Acids Research. 2004. 32 (5): 1792–97

17. Eddy S.R. Profile hidden Markov models // Bioinformatics. 1998. 14 (9): 755–63

18. Eddy S.R., Pearson, William R. Accelerated Profile HMM Searches // PLoS Computational Biology. 2011. 7 (10)

19. Eickbush T.H., Moudrianakis E.N. The histone core complex: an octamer assembled by two sets of protein-protein interactions // Biochemistry. 1978. 17(23): 4955–4964.

20. Elsaesser S.J., Goldberg A.D., Allis C.D. New functions for an old variant: no substitute for histone H3.3 // Curr. Opin. Genet. Dev. 2010. 20(2): 110–117.

21. Elsässer S.J., Noh K.-M., Diaz N., Allis C.D., Banaszynski L.A. Histone H3.3 is required for endogenous retroviral element silencing in embryonic stem cells // Nature. 2015. 522(7555): 240–244.

22. Fischle W., Tseng B.S., Dormann H.L., Ueberheide B.M., Garcia B.A., Shabanowitz J., Hunt D.F., Funabiki H., Allis C.D. Regulation of HP1–chromatin binding by histone H3 methylation and phosphorylation // Nature. 2005. 438(7071): 1116–1122.

23. Gehre M., Bunina D., Sidoli S., Lübke M.J., Diaz N., Trovato M., Garcia B.A., Zaugg J.B., Noh K.-M. Lysine 4 of histone H3.3 is required for embryonic stem cell differentiation, histone enrichment at regulatory regions and transcription accuracy // Nat. Genet. 2020. 52(3): 273–282.

24. John N. Reeve, Kathleen Sandman, Charles J. Daniels. Archaeal Histones, Nucleosomes, and Transcription Initiation // Cell, 1997.89, 999–1002.

25. Kamakaka R.T., Biggins S. Histone variants: deviants? // Genes Dev. 2005. 19(3): 295–316.

26. Kasinsky H.E., Lewis J.D., Dacks J.B., Ausio J. Origin of H1 linker histones. // FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol. 2001. 15(1): 34–42.

27. Khochbin S. Histone H1 diversity: bridging regulatory signals to linker histone function // Gene. 2001. 271(1): 1–12.

28. Kiefer C.M., Hou C., Little J.A., Dean A. Epigenetics of beta-globin gene regulation. // Mutat. Res. 2008. 647(1–2): 68–76.

29. Kimura M. The neutral theory of molecular evolution: A review of recent evidence // Jpn J Genet. 1991. 66 (4): 367-86.

30. Lyons S.M., Cunningham C.H., Welch J.D., Groh B., Guo A.Y., Wei B., Whitfield M.L., Xiong Y., Marzluff W.F. A subset of replication-dependent histone mRNAs are expressed as polyadenylated RNAs in terminally differentiated tissues. // Nucleic Acids Res. 2016. 44(19): 9190–9205.

31. Makalowska I., Ferlanti E.S., Baxevanis A.D., Landsman D. Histone Sequence Database: sequences, structures, post-translational modifications and genetic loci // Nucleic Acids Res. 1999. 27(1): 323–324.

32. Mariño-Ramírez L., Hsu B., Baxevanis A.D., Landsman D. The Histone Database: a comprehensive resource for histones and histone fold-containing proteins. // Proteins. 2006. 62(4) 838–842.

33. Mariño-Ramírez L., Levine K.M., Morales M., Zhang S., Moreland R.T., Baxevanis A.D., Landsman D. The Histone Database: an integrated resource for histones and histone fold-containing proteins // Database (Oxford). 2011. 2011, bar048–bar048.

34. Margulis, L., Schwartz, K.V. Five Kingdoms: an illustrated guide to the Phyla of life on earth. 3rd edition // Freeman: New York, NY (USA). 1998. 520.

35. Marzluff W.F., Wagner E.J., Duronio R.J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail // Nat. Rev. Genet. 2008. 9(11): 843–854.

36. Oudet P., Gross-Bellard M., Chambon P. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit // Cell. 1975. 4(4): 281–300.

37. Öztürk M.A., De M., Cojocaru V., Wade R.C. Chromatosome Structure and Dynamics from Molecular Simulations // Annu. Rev. Phys. Chem. 2020. 71(1): 101–119.

38. Palozola K.C., Donahue G., Liu H., Grant G.R., Becker J.S., Cote A., Yu H., Raj A., Zaret K.S. Mitotic transcription and waves of gene reactivation during mitotic exit // Science. 2017. 358(6359): 119–122.

39. Pehrson J.R., Fried V.A. MacroH2A, a core histone containing a large nonhistone region. // Science. 1992. 257(5075): 1398–1400.

40. Pinto D.M.S., Flaus A. Structure and function of histone H2AX. // Subcell. Biochem. 2010. 50, 55–78.

41. Sharma A.B., Dimitrov S., Hamiche A., Van Dyck, E. Centromeric and ectopic assembly of CENP-A chromatin in health and cancer: old marks and new tracks. // Nucleic Acids Res. 2019. 47(3): 1051–1069.

42. Simpson R.T. Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones // Biochemistry. 1978. 17(25): 5524–5531.

43. Singh R., Bassett E., Chakravarti A., Parthun M.R. Replication-dependent histone isoforms: a new source of complexity in chromatin structure and function // Nucleic Acids Res. 2018. 46(17): 8665–8678.

44. Simpson, Alastair G.B., Roger, A.J. The real "kingdoms" of eukaryotes // Current Biology. 2005. 14 (17): 693–696

45. Skene P.J., Henikoff S. Histone variants in pluripotency and disease. // Development. 2013. 140(12): 2513–2524.

46. Sullivan S.A., Aravind L., Makalowska I., Baxevanis A.D., Landsman D. The histone database: a comprehensive WWW resource for histones and histone fold-containing proteins. // Nucleic Acids Res. 2000. 28(1) 320–322.

47. Sullivan S., Sink D.W., Trout K.L., Makalowska I., Taylor P.M., Baxevanis A.D., Landsman D. The Histone Database // Nucleic Acids Res. 2002. 30(1): 341–342.

48. Suto R.K., Clarkson M.J., Tremethick D.J., Luger K. Crystal structure of a nucleosome core particle containing the variant histone H2A.Z // Nat. Struct. Biol. 2000. 7(12): 1121–1124.

49. Tagami H., Ray-Gallet, D., Almouzni G., Nakatani Y. Histone H3.1 and H3.3 complexes mediate nucleosome assembly pathways dependent or independent of DNA synthesis. // Cell. 2004. 116(1): 51–61.

50. Talbert P.B., Ahmad K., Almouzni G., Ausió J., Berger F., Bhalla P.L., Bonner W.M., Cande W.Z., Chadwick B.P., Chan S.W.L., Cross G.A.M., Cui L., Dimitrov S.I., Doenecke D., Eirin-López J.M., Gorovsky M.A., Hake S.B., Hamkalo B.A.,

Holec S., Jacobsen S.E., Kamieniarz K., Khochbin S., Ladurner A.G., Landsman D., Latham J.A., Loppin B., Malik H.S., Marzluff W.F., Pehrson J.R., Postberg J., Schneider R., Singh M.B., Smith M.M., Thompson E., Torres-Padilla M.-E., Tremethick D.J., Turner B.M., Waterborg J.H., Wollmann H., Yelagandula R., Zhu B., Henikoff S. A unified phylogeny-based nomenclature for histone variants // Epigenetics Chromatin. 2012. 5(1): 7.

51. Talbert P.B., Henikoff S. Histone variants on the move: substrates for chromatin dynamics // Nat. Rev. Mol. Cell Biol. 2017. 18(2): 115–126.

52. Talbert P.B., Henikoff S. What makes a centromere? // Exp. Cell Res. 2020. 389(2): 111895.

53. Talbert P.B., Henikoff S. Phylogeny as the basis for naming histones. // Trends Genet. 2013. 29(9): 499–500.

54. Talbert P.B., Meers M.P., Henikoff S. Old cogs, new tricks: the evolution of gene expression in a chromatin context // Nat. Rev. Genet. 2019. 20(5): 283–297.

55. Thomas J.O., Kornberg R.D. An octamer of histones in chromatin and free in solution // Proc. Natl. Acad. Sci. 1975. 72(7): 2626–2630.

56. Tremethick D.J. Higher-Order Structures of Chromatin: The Elusive 30 nm Fiber // Cell. 2007. 128(4): 651–654.

57. Turner B.M. Reading signals on the nucleosome with a new nomenclature for modified histones. // Nat. Struct. Mol. Biol. 2005. 12(2): 110–112.

58. Udugama M., M Chang F.T., Chan F.L., Tang M.C., Pickett H.A., R McGhie, J.D., Mayne L., Collas P., Mann J.R., Wong L.H. Histone variant H3.3 provides the heterochromatic H3 lysine 9 tri-methylation mark at telomeres // Nucleic Acids Res. 2015. 43(21): 10227–10237.

59. Watson J.D., Crick F.H.C. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid // Nature. 1953. 171,737.

60. Woese C.R., Kandler O., Wheelis M.L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eukaryota // Proceedings of the National Academy of Sciences of the USA. 1990. 87 (12): 4576–9.

61. Zhao Z., Shilatifard A. Epigenetic modifications of histones in cancer // Genome Biol. 2019. 20(1): 245.
62. Zhou B.-R., Jiang J., Feng H., Ghirlando R., Xiao T.S., Bai Y. Structural Mechanisms of Nucleosome Recognition by Linker Histones // Mol. Cell. 2015.