Приложение 1

К статье А.И. Дергилева, А.М. Спициной, И.В. Чадаевой, А.В. Свичкарева, Ф.М. Науменко, Е.В. Кулаковой, Э.Р. Галиевой, Е.Е. Витяева, М. Чен, Ю.Л. Орлова «Компьютерный анализ совместной локализации сайтов связывания транскрипционных факторов в геноме по данным ChIP-seq»

Разработана компьютерная программа расчета кластеров сайтов связывания различных транскрипционных факторов по данным геномных координат пиков ChIP-seq. Рассмотрены статистические особенности распределения сайтов связывания транскрипционных факторов в геноме мыши, полученных с помощью ChIP-seq экспериментов в эмбриональных стволовых клетках. Определены кластеры сайтов, содержащие четыре и более сайта связывания различных транскрипционных факторов в геноме мыши, описано их расположение относительно регуляторных районов генов. Подтверждено присутствие двух типов совместной локализации сайтов: кластеры, содержащие сайты связывания факторов Oct4, Nanog, Sox2, расположенные в дистальных районах, и кластеры, содержащие сайты связывания n-Myc, c-Myc, расположенные в основном в промоторных районах генов мыши. Анализ новых данных ChIP-seq по связыванию транскрипционных факторов Nr5a2, Tbx3 в том же типе клеток подтвердил разделение кластеров сайтов связывания транскрипционных факторов два типа – содержащие и не содержащие сайты связывания регуляторов плюрипотентности (Oct4, Nanog и Sox2). Разработана компьютерная программа статистической обработки данных расположения генов для анализа экспериментальных данных локализации сайтов, полученных методами ChIP-seq в геномах мыши и человека. С помощью этой программы выявлено наличие предпочтений в положении сайтов связывания транскрипционных факторов различных типов. Рассчитаны расстояния между ближайшими сайтами связывания ТФ группы Oct4, Nanog, Sox2 и сайтами связывания других факторов в кластерах сайтов, которые служат основой для анализа совместного связывания белковых комплексов с ДНК. Оценена доля присутствия нуклеотидных мотивов сайтов связывания транскрипционных факторов в геномных участках ChIP-seq; уточнены известные нуклеотидные мотивы. Показана корреляция присутствия мотивов с интенсивностью связывания ChIP-seq. Программы, реализующие разработанные компьютерные методы оценки кластеризации сайтов связывания различных транскрипционных факторов для новых данных ChIP-seq, доступны по запросу к авторам.

Дополнительные материалы 1. Колокализация сайтов связывания транскрипционных факторов (таблица корреляций)

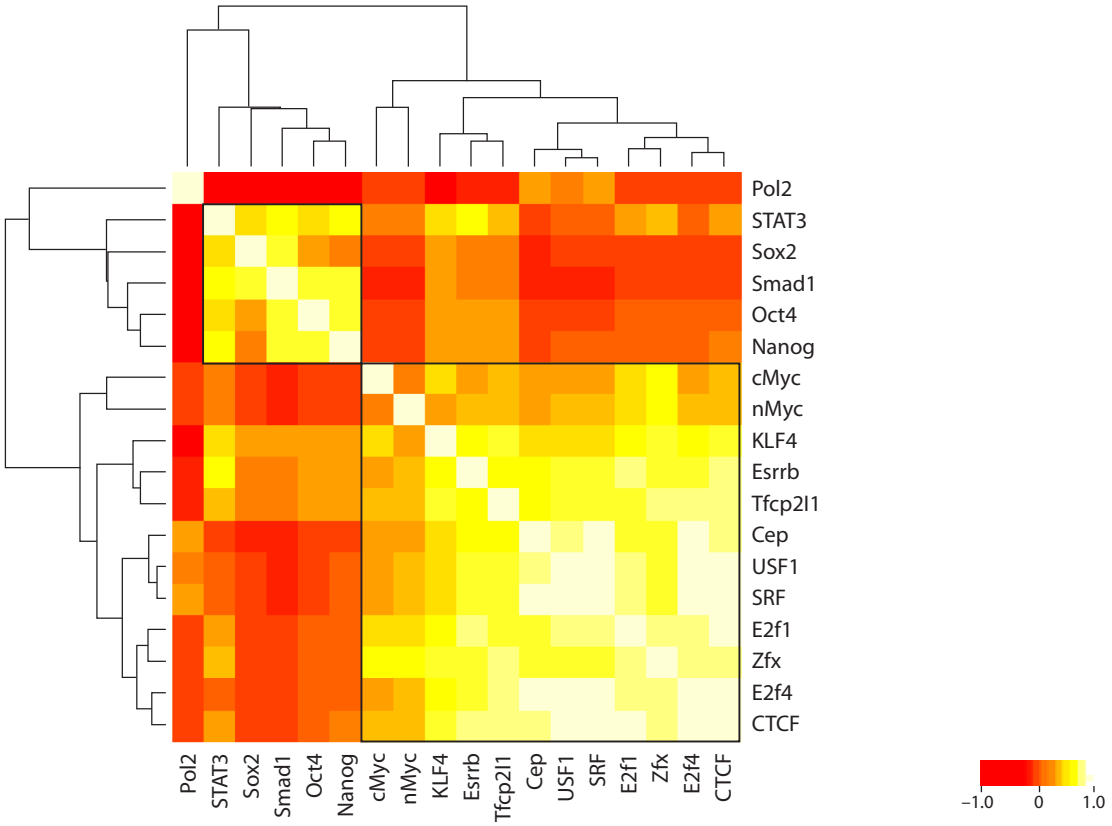


Рис. 1. Тепловая карта для набора из 18 ТФ.

Из рис. 1 видно, что среди 18 факторов Nanog, Sox2, Oct4, Smad1, и STAT3 имеют тенденцию встречаться совместно более часто, также выделяется вторая группа, состоящая из факторов n-Myc, c-Myc, E2f1 и Zfx.

Для сравнения была построена тепловая карта для набора из 10 факторов – E2f1, USF1, KLF4, Sox2, Smad1, STAT3, c-Myc, Tfcp2l1, Zfx, Oct4 (факторы были выбраны случайным образом из имеющегося списка).

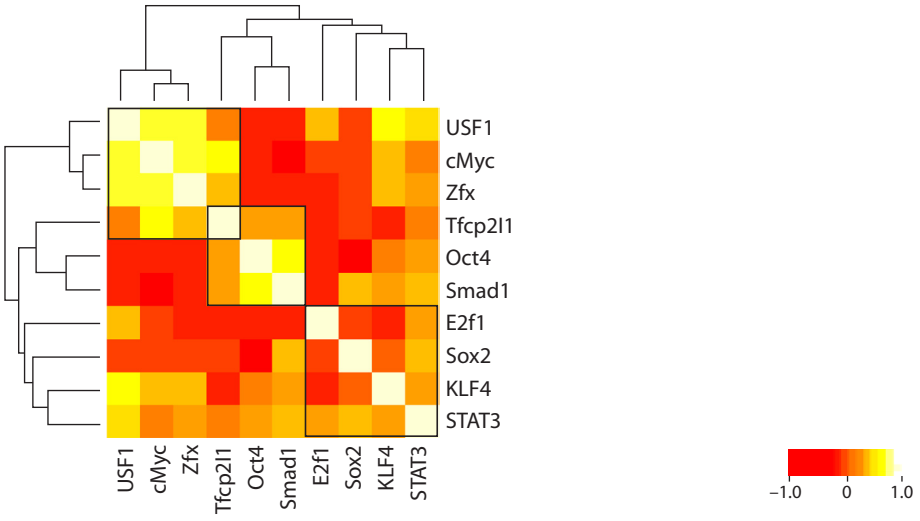


Рис. 2. Тепловая карта для набора из 10 ТФ.

Среди 10 выбранных факторов USF и c-Myc имеют тенденцию встречаться совместно более часто, остальные группы часто встречаемых факторов очень разнятся.

Проведен анализ числа кластеров в геноме зависимости от расположения (промоторные и дистальные сайты).

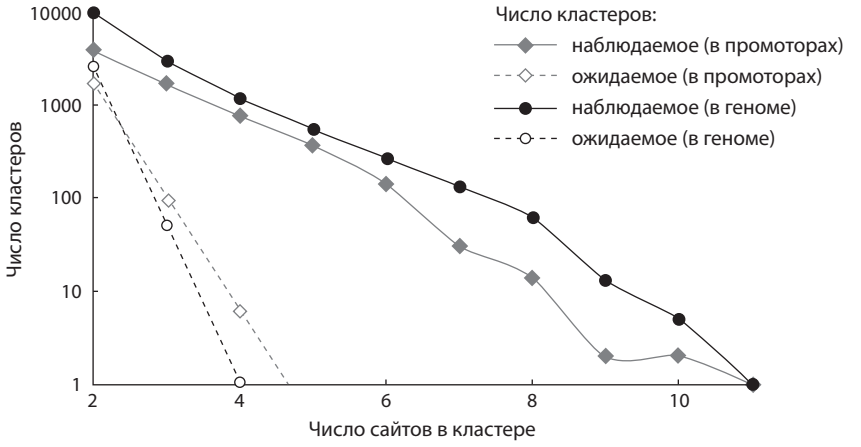


Рис. 3. Наблюдаемые значения числа кластеров в зависимости от числа сайтов в кластере и оценки получения кластеров по случайным причинам из того же числа сайтов в тех же районах генома (компьютерная симуляция).

Дополнительные материалы 2. Статистика кластеров

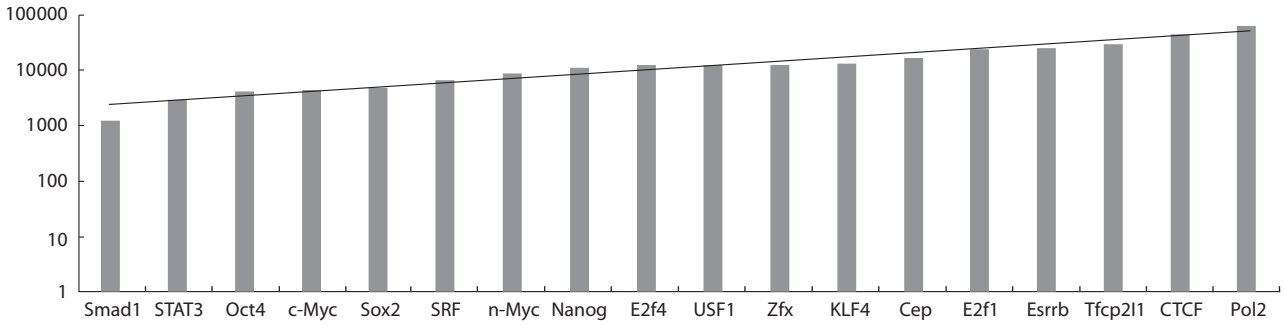


Рис. 1. Статистика встречаемости среди 18 ТФ.

Показано, что чаще встречается Pol2, в то время как Smad1 встречается реже. Наиболее распространенными являются одиночные кластеры, а менее распространенными – кластеры с наибольшим числом сайтов. Данные получены при параметре зазора равным 200 нуклеотидов, что соответствует точности определения пиков ChIP-seq.

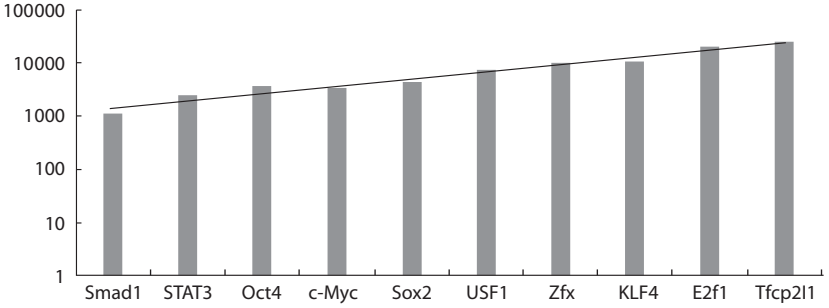


Рис. 2. Статистика встречаемости среди 10 ТФ.

Для исследования распределения кластеров сайтов связывания в геноме был проведен дополнительный статистический анализ для набора из 10 факторов – E2f1, USF1, KLF4, Sox2, Smad1, STAT3, c-Myc, Tfcp2l1,Zfx, Oct4 (факторы были выбраны случайным образом из имеющегося списка для тестирования программы).

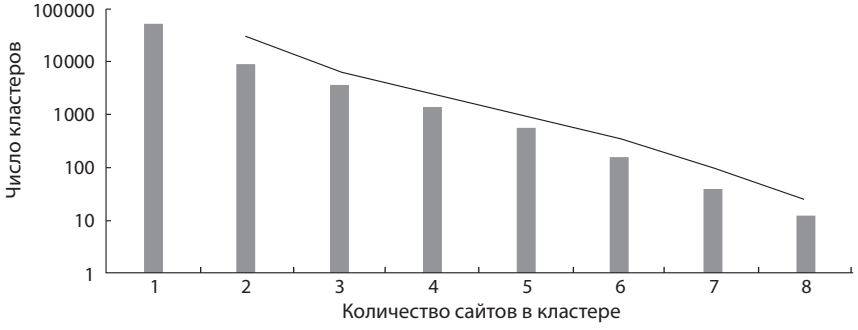
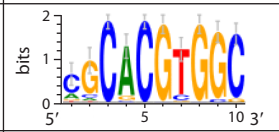
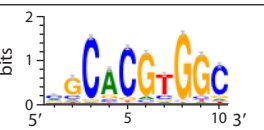
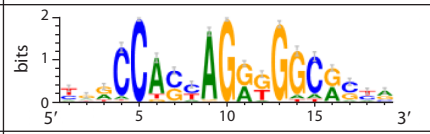
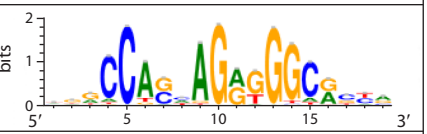
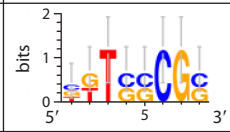
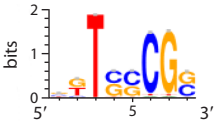
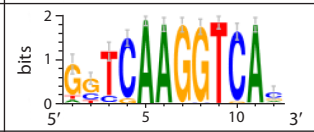
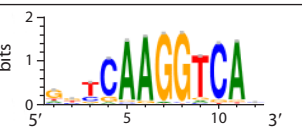
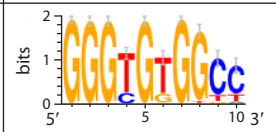
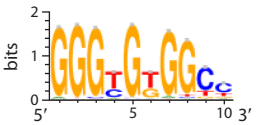
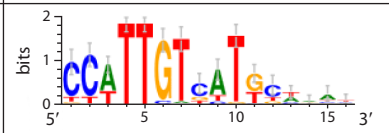
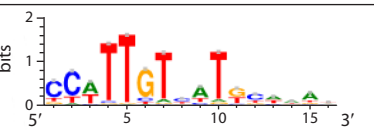
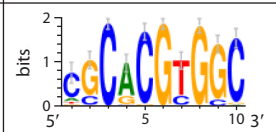
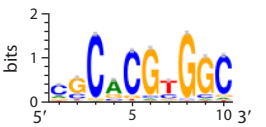
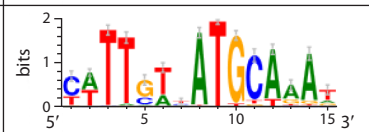
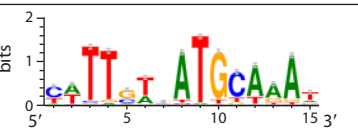
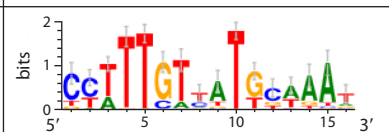
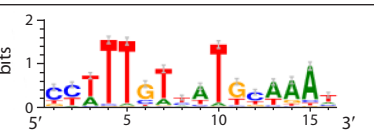
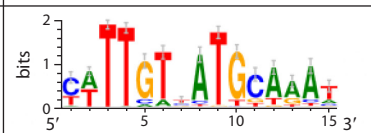
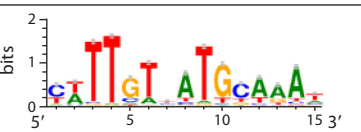
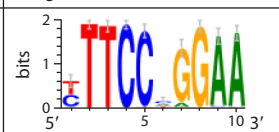
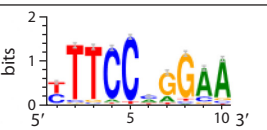
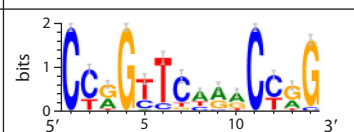
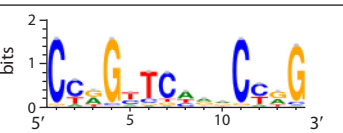
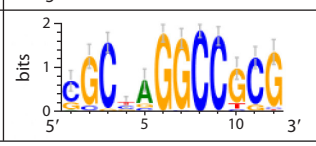
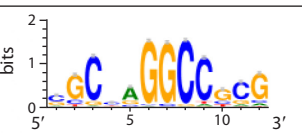


Рис. 3. Зависимость размера кластеров от числа сайтов среди 10 ТФ.

Из первой гистограммы (рис. 2) видно, что чаще всего встречается Tfcp2l1, в то время как Smad1 встречается реже, т. е. Smad1 по-прежнему встречается реже остальных.

Из следующей гистограммы (рис. 3), представляющей зависимость размера кластеров от числа сайтов среди 10 ТФ, видно, что наиболее распространенными являются одиночные кластеры, как и ранее, а менее распространенными – кластеры с наибольшим числом сайтов (8).

Дополнительные материалы 3. Сравнение лого транскрипционных факторов

ТФ	Лого оригинальных мотивов связывания ТФ с ДНК TF (TRANSFAC)	Лого уточненных мотивов (после ChIP-seq)
c-Myc		
CTCF		
E2f1		
Esrrb		
K1f4		
Nanog		
n-Myc		
Oct4		
Smad1		
Sox2		
STAT3		
Tcfcp2i1		
Zfx		

Дополнительные материалы 4

Пересчет мотива по геномной ДНК по совпадениям, найденным с помощью весовой матрицы, выполнен для данных по геномам мыши и человека. Для сайтов связывания CTCF видна высокая консервативность мотива.

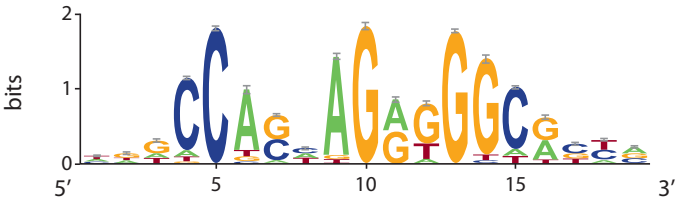
Геном	Лого CTCF
Mus musculis (mm10)	
Homo sapiens (hg19)	

Рис. 1. Анализ нуклеотидных мотивов в пиках ChIP-seq в геномах различных организмов показывает высокую консервативность сайтов связывания CTCF.

Выбор порога распознавания мотива может уточняться. Оптимальным значением порога, минимизирующим ошибки первого и второго рода, выбрано 0.8. На рис.2 показан процент присутствия мотивов сайтов CTCF в зависимости от уровня порога распознавания весовой матрицы в нуклеотидных последовательностях пиков ChIP-seq и в случайных геномных последовательностях того же размера, что подтверждает оптимальность такого выбора порога.

Выполнена качественная оценка присутствия мотивов в зависимости от интенсивности связывания (высоты пиков). Пики ChIP-seq были сортированы по высоте (интенсивности связывания); полученный ранжированный список разбит на квартили. В каждой квартили подсчитаны число и относительная доля найденных мотивов. Показана положительная ассоциация присутствия мотивов (одного или более одного мотива в участке) в зависимости от квартили (25% элементов упорядоченного списка).

Из рис. 3 видна зависимость присутствия мотива от интенсивности связывания ДНК с белком в эксперименте ChIP-seq.

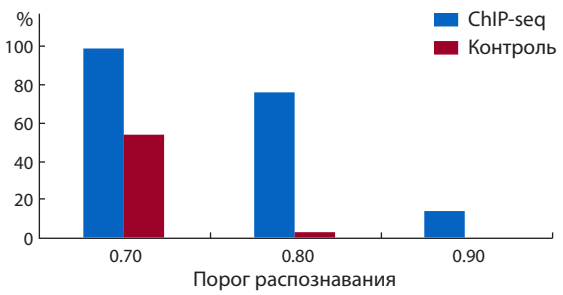


Рис. 2. Процент распознавания (присутствия) нуклеотидных мотивов сайтов CTCF в участках пиков ChIP-seq и контроле в зависимости от порога распознавания (0.7–0.9).

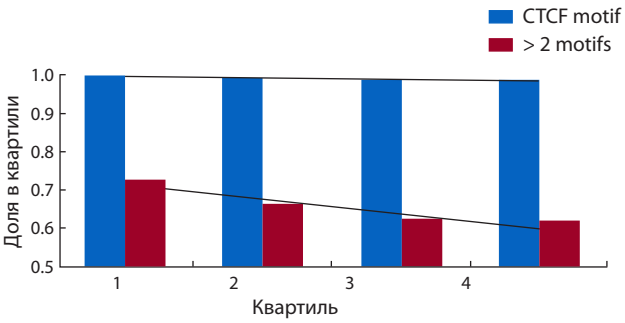


Рис. 3. Доля присутствия мотива сайтов CTCF в пиках ChIP-seq в зависимости от интенсивности связывания (ранжирования по четырем квартилям, 1 – наибольшая интенсивность, 4 – наименьшая).